VALIDATING POLICY RATINGS: THE SUBSTANTIVE ASPECT OF CONSTRUCT VALIDITY FOR RATINGS OF SCHOOL TOBACCO POLICIES

by

Cornelia Zeisser

B.A. (Hons.), Simon Fraser University, 2002
M.A. (Hons.), The University of British Columbia, 2004

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Measurement, Evaluation and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2010

ABSTRACT

This dissertation investigated the substantive aspect of construct validity in the context of Canadian school tobacco policy ratings. The objective was to provide a better understanding of score meaning via the process of expert rater responding while rating school tobacco policies. Study one described Canadian school tobacco policies and identified policy characteristics. Written tobacco policies (N=196) were obtained from schools and boards across 10 Canadian provinces that participated in the Youth Smoking Survey. Policies were coded to identify characteristics associated with effectiveness in preventing student tobacco use. Smoking prevention education and cessation access were identified as key policy components that need to be addressed more strongly. Policy characteristics identified in study one formed the basis for study two. The objective of study two was to examine the cognitive processes that generate raters' responses, identify rating obstacles and how raters overcome them. A think-aloud protocol was conducted with two expert tobacco policy raters who rated 12 tobacco policies using the Stephens & English rubric. Policies were sampled to reflect characteristics (type, length and comprehensiveness) identified in study one. Transcripts were coded to identify super-categories (rater behaviors), main categories (major cognitive processes at the item level) and subcategories to describe main processes in more detail. Categories and their interrelationships, rating obstacles and raters' coping strategies are presented and a series of cognitive process models of rating is proposed. Findings suggest that raters use similar main processes explainable by similar sub-processes regardless of policy type rated. There was variation in rating obstacles and rater coping when different policy types were rated. The cognitive process models contribute to the substantive aspect of construct validity by providing explanations for score variation and enhancing understanding of score meaning. Explanation is sufficient when policies are comprehensive but is limited if based on short, less comprehensive policies. Implications for practice and policy recommendations are discussed.

TABLE OF CONTENTS

ABSTRACT ii

TABLE OF CONTENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

ACKNOWLEDGEMENTS ix

DEDICATION x

CO-AUTHORSHIP STATEMENT xi

CHAPTER ONE: INTRODUCTION TO THE DISSERTATION 1

General Introduction to the Research Problem 1

Methodological Review 3

The Changing View of Validity 3

The Substantive Aspect of Construct Validity 5

Sources of Validity Evidence and Reporting of Validity Evidence 6

Importance of Gathering Validity Evidence by Examining Processes of Responding 7

Dissertation Objectives 10

Structure of the Dissertation 15

References 18

CHAPTER TWO: A DESCRIPTIVE AND COMPARATIVE ANALYSIS OF CANADIAN SCHOOL TOBACCO CONTROL POLICIES 23

Introduction and Literature Review 23

Methods 26

Results 30

Discussion 34

References 37

CHAPTER THREE: AN INVESTIGATION INTO SCHOOL TOBACCO POLICY RATING FROM THE PERSPECTIVE OF THE SUBSTANTIVE ASPECT OF CONSTRUCT VALIDITY 40

Introduction and Literature Review 40

Changing Approaches to Gathering Validity Evidence 44

The Substantive Aspect of Construct Validity 44

The Status of Reporting on Cognitive Processes as Sources of Validity Evidence 45

Importance of Studying Response Processes Above and Beyond Traditional Psychometrics 46

Validity of Inferences from Rating Data 49

The Think-Aloud Method 54

Potential Limitations of the TA Method 55

Distinction of the TA Method from Introspection 57

Methodological Significance of TA Data 58

Protocol Analysis and the Specific Use of TA for Rating Tasks 60

Methods 62

Analysis 83

Results 89

Discussion 113

References 134

CHAPTER FOUR: CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS 146

Review of the Purpose of the Dissertation 146

Summary of Findings in Light of Current Research in the Field 146

Study One Findings 146

Study Two Findings 149

Importance of the Linkage Between Studies One and Two 151

Variations in the Rating Processes as a Function of Policy Characteristics 151

Implications for Research and Practice 155

Future Research Directions and Recommendations 156

Contributions of the Dissertation 158

References 162

APPENDIX A 165

LIST OF TABLES

Table 1 Variables and Coding Used in Characterizing Canadian School Tobacco Control Policies…………………………………………………………………………………...28

Table 2 Think-Aloud Study Policy Sampling Frame and Order in Which Raters Coded Policies…………………………………………………………………………………...68

Table 3 Items with Inter-Rater Agreement of < 90 % Used by Both Raters…………………….70

Table 4 Items with High (90-95%) Inter-Rater Agreement Randomly Assigned To Rater A or B……………………………………………………………………………………………….71

Table 5 Subcategories Description and Relation to Super – Categories………………………...91

Table 6 Main Cognitive Processes in Raters: Categories at the Level of the Policy Rating Item with Total Frequencies……………………………………………………………………….96

Table 7 Subcategories of Raters' Cognitive Processes at the Unit of Thought-Level with Excerpt Examples and Total Frequencies……………………………………………………………….98

Table 8 Main Cognitive Processes in Raters: Categories with Total Frequencies Broken Down by the Policy Sampling Frame………………………………………………………………..100

LIST OF FIGURES

Figure 1. Distribution of Tobacco Control Policy Characteristics. 31

Figure 2. Summary Analysis: Percentage of Tobacco Control Policies Addressing Key Policy Components. 33

Figure 3. Information Provided To Raters Prior To the Think-Aloud Task. 75

Figure 4. Flow Diagram of Data Analysis. 88

Figure 5. Diagram Demonstrating Category Structure Using Examples of Categories. 93

Figure 6. The Cognitive Process Model of Policy Rating For all Policies. 107

Figure 7. The Cognitive Process Model of Policy Rating For Board –Developed Policies. 108

Figure 8. The Cognitive Process Model of Policy Rating For School –Developed Policies...... 109

ACKNOWLEDGEMENTS

I express my sincere gratitude to the faculty members who have guided and supported me in this long process, in particular my committee: Professor Bruno Zumbo, Dr. Chris Lovato and Dr. Kim Schonert-Reichl, as well as Dr. Richard Young. This work would have been impossible without their support. I owe my special thanks to Professor Bruno Zumbo, whose continued commitment and support enabled me to complete this journey whilst enjoying it, and Dr. Chris Lovato, whose work provided inspiration and a foundation for my dissertation. Special thanks are also owed to my colleagues in the Tobacco Research Unit at the University of British Columbia, Brad Lowen, B.A. and Allison Pullman, MSc., for offering their expertise in tobacco policy rating.

I fondly thank Dr. Barry Beyerstein, who passed away in 2007. His mentorship and inspiration in earlier years helped me find this path.

Last but not least, I thank my fellow students who gave me inspiration and encouragement.

Part of this research was supported by a Canadian Institute of Health Research (CIHR) Strategic Training Program in Tobacco Research (STPTR) grant (2008).

DEDICATION

To Brian, for his loving and unconditional encouragement, support and patience

To my parents, to whom I will be forever thankful

To my friends who cheered me on along the path and sometimes kept me on it, in particular a wise woman Galina with her words during a time rough and treacherous: "You defend PhD once, but PhD defends you all your life"!

CO-AUTHORSHIP STATEMENT

Study one was co-authored by Cornelia Zeisser, Chris Y. Lovato, Bruno D. Zumbo, Allison Pullman and Steve Manske. As a first author, I was in charge of formulating the study, performing the research and literature review, analyzing the data and preparing the manuscript. The first three authors contributed jointly to the identification of the research design.

Study two was co-authored by Cornelia Zeisser, Bruno D. Zumbo, Chris Y. Lovato and Richard Young. As the first author, I was involved with formulating the study, designing the study, performing the research and literature review, data analysis and preparation of the manuscript. The first two authors jointly contributed to the identification of the research project. Dr. Richard Young provided significant advice and input with the data analysis.

CHAPTER ONE: INTRODUCTION TO THE DISSERTATION

General Introduction to the Research Problem

Researchers in health promotion, education and social science desire to model and predict outcomes from large community interventions such as policies. For this purpose, the task of rating policies has become an important research activity for generating policy scores that one can use as predictors of outcomes. That is, researchers perform ratings to quantify policy strength. The process of policy rating is relevant in many areas, such as in education, using anti-bullying policies (Ordonez, 2007) or health promotion, using school drug policy (Evans-Whipp, Beyers, Lloyd, Lafazia, Toumbourou, Arthur et al., 2004). Policy rating is also applied in clean-air bylaw coding (Nykiforuk, 2004). The specific focus of this dissertation is on policy rating in the context of school tobacco control (Lovato, Sabiston, Hadd, Nykiforuk & Campbell, 2007; Murnaghan, Sihvonen, Leatherdale & Kekki, 2007; Adams, Jason, Pokorny & Hunt, 2009; Boyce, Mueller, Hogan-Watts & Luke, 2009; Tompkins, Dino, Zedowsky, Harman & Shaler, 1999; Moore, Roberts & Tudor-Smith, 2001). However, this dissertation examined policy rating as a process in general, as it would be applicable to any other research area. Such an examination of policy rating in general is, to the best of our knowledge, not yet available in the literature.

Researchers apply ratings to quantify policy variables of interest; hence, the value of policy rating lies in enabling researchers to explore the connection between the policy and outcomes of interest. For example, tobacco prevention researchers aim to use policy scores as explanatory variables; hence, characterizing and quantifying policies are important activities for today's researchers.

In order to quantify a policy, it is important to define clearly what a policy is. However, the majority of studies on policy do not provide a clear definition of a policy. Generally, policy is defined as formal laws, rules and regulations enacted by elected officials, as well as organizational guidelines, deliberate plans of action and agency decisions to guide social norms and achieve specific outcomes (Milio, 2001; Schmid, Pratt & Howze, 1995). More specifically, in the context of school tobacco prevention, a policy is a formal statement of principles and rules established by the school or school board to provide guidance regarding the implementation and enforcement of no-smoking rules. In this dissertation, a written school tobacco policy is defined as any document – in hard copy or available online – that describes a school's rules and regulations regarding students' and teachers' smoking, possession of tobacco products, smoke-free environments, locations of smoking restrictions and consequences of violations for various groups (e.g., students, teachers, school visitors). The school tobacco policy document can be in a variety of forms, such as stand-alone document, a part of a student/parent handbook, or a code of conduct. The tobacco policy of a school can also be a policy developed by the local school board and adopted by the school as their own tobacco policy.

A rating can be broadly defined as a classification according to order or grade. In the context of this dissertation, rating is defined as the classification of tobacco policy content components as either present or absent. Specifically, tobacco policy rating is defined as expert raters assigning a numeric value on various content dimensions using a coding rubric to generate a total score for each policy to reflect its strength.

In the endeavour of policy quantification, researchers often face the challenge to generate policy indicators from meaningful scores, i.e., scores from which they can draw valid inferences. Approaches to this challenge tend to be limited to using traditional psychometrics such as correlating a score with another criterion in order to establish to what degree one can draw valid inferences from the score (Hogan & Agnello, 2004). Frequently, studies only report on the inter-rater reliability of ratings (Tompkins, Dino, Zedosky, Harman & Shaler, 1999; Stephens & English, 2002; Adams, Jason, Pokorny & Hunt, 2009). However, these approaches are limited because they do not provide investigators with information about the nature and meaning of the scores. To fill this gap, this dissertation is an investigation into policy score meaning via a novel approach to validity that aims to provide a better understanding of policy score meaning by illuminating the process of how expert raters generate a policy rating score. That is, the need for this dissertation arose from the fact that researchers frequently use policy ratings and the challenge to create policy scores from which one can draw inferences that are valid and justified.

Methodological Review

The Changing View of Validity

The importance of validity has been widely acknowledged in the social and health sciences. Without validation, any inferences from a score or measure are of limited use since they may be meaningless and inappropriate (Hubley & Zumbo, 1996; Zumbo, 2007). However, the concept of validity and the process of validation have continued to change throughout the 20th century since their inception (Hubley & Zumbo, 1996). The early thinking (early – to mid 1900s) was centered around the criterion-based approach to validity. This approach was taken by Anastasi (1950) and focused on the test as a means to predict future outcomes and behaviours based on observable criteria. During the 1950s, the focus shifted from observables to unobservables, when Cronbach & Meehl (1955) introduced the construct model of validity. These authors sought to support meaningfulness via the nomological network showing how scores from a test reflect the underlying theoretical constructs. Loevinger (1957) pointed out the challenges and threats to construct validity, namely construct underrepresentation and construct-irrelevant variance. This thinking was strongly influenced by the field of psychology and modeled after theories of learning and behaviour. With the cognitive revolution in psychology during the 1960s and 1970s, the construct validity model was the prevalent model. Building on these foundations, Messick introduced the issues of consequences and test interpretation (1975, 1980, 1988, 1989, 1995, and 1998). Kane (2001) argued that there are strong and weak forms of construct validity. The weak form is characterized by a correlation of test scores with other variables to serve as evidence. The challenge with this approach is that with the weak form of construct validity, a test has as many validities as it can have correlations with other variables. On the other hand, the strong form of construct validity is founded in well-articulated models and theories and well-designed empirical tests of the theory.

Zumbo (2005, 2007) argues that the strong form of construct validity should provide an explanation for the test scores. That is, the theory should produce an explanation for the observed variation in test scores. Building on this argument, Zumbo (2009) further emphasizes that, while validity is a matter of inference and weighing of evidence, explanatory consideration should guide the inferences one makes from scores. Having explanation as a regulative ideal, validity is the explanation of score variation, and validation is the process of developing and testing the explanation. Further, Zumbo (2009) argues that understanding and explanation – central to the notion of validity – arise from carefully balancing contrastive data and competing views. Specifically, understanding why an individual responded a certain way to an item or scored a particular value on a scale would go a long way toward bridging the inferential gap between test scores and constructs (Zumbo, 2009). According to this view, evidence for validity is only established when one has developed an explanatory model of the variation in item responses and scale scores and the variables mediating, moderating or otherwise influencing the response. The next section provides more detail on Messick's validity view, with particular focus on the substantive aspect of construct validity.

The Substantive Aspect of Construct Validity

Messick (1994, 1995) discusses six aspects of construct validity: content, substantive, structural, generalizability, external and consequential. The content aspect of construct validity includes evidence of content relevance, representativeness and technical quality. The structural aspect appraises fidelity of the scoring structure to the structure of the construct domain of interest. The generalizability aspect examines the degree to which score interpretations generalize across populations, tasks and settings. The external aspect includes convergent and discriminant evidence from multitrait-multimethod studies as well as evidence of criterion relevance and applied utility. The consequential aspect appraises value implications of score interpretation as a basis for action and the actual and potential consequences of test use in regard to fairness and bias. The substantive aspect of construct validity refers to theoretical rationales for the observed consistencies in responses and includes cognitive process models (Messick, 1994). This aspect focuses on the need for empirical evidence of response consistencies reflective of domain processes. As such, the substantive aspect, instead of relying on traditional psychometrics such as correlations or factor analyses, emphasizes the role of substantive theories and process modeling in identifying domain processes to be revealed in assessment tasks. In the substantive aspect of construct validity, Messick (1994, 1995) suggests the use of theory and process models as important, albeit rarely explored avenues for gathering evidence in support of construct validity. The substantive aspect focuses on the role of substantive theories and process modeling in identifying domain processes to be revealed in assessment tasks (Messick, 1989, 1994 and 1995). Hence, this aspect emphasizes the need for empirical evidence of response consistencies reflective of domain processes. This evidence could be obtained via a think-aloud (TA) protocol during task performance.

In the substantive aspect of construct validity, two important points are involved. One is the need for tasks that appropriately sample domain processes; the other is the need to move beyond traditional professional judgment of content to gather empirical evidence that the sampled processes are indeed engaged in by respondents during the assessment task. Hence, the substantive aspect adds the need for the process representation of the construct of interest and the degree to which these processes are reflected in construct measurement (Messick, 1995). The following section provides some background on the sources of validity evidence and the actual status of reporting this evidence in published research.

Sources of Validity Evidence and Reporting of Validity Evidence

The literature reveals that some sources of validity evidence are essentially ignored in validity reports (Cizek, Rosenberg & Koons, 2008). For example, Hogan & Agnello (2004) found that “only 55% of the reports included any type of validity evidence” (p. 802). Further, the vast majority reported only correlations with other variables as a source of validity evidence. These findings were based on the authors’ investigation of validity reporting practices in a sample of 696 research reports listed in the APA’s Directory of Unpublished Experimental Mental Measures (Goldman & Mitchell, 2003). In a similar vein, Cizek et al. (2008) found that the majority of articles they investigated reported only on construct validity (58.0%), criterion-related/concurrent validity (50.9%) and some on content validity (48.4%). Only 1.8% of articles reported on response processes as a type of validity evidence. Further, the reporting on response processes as a source of validity was found only for developmental tests (5.9%), behavioural (4.0%), achievement tests (3.7%) and cognitive skills tests (1.5%); no reports about response processes were found for studies involving raters performing the task of coding written documentation such as policies. It appears that the latter type of study would lend itself particularly well to an exploration of rater response processes as a source of validity evidence, since a study of rater responding can be easily implemented during a rating task and can provide valuable clues and rich information as to how raters generate their codes during the task. The following section briefly discusses the newer approach to gathering validity evidence via the process of responding.

Importance of Gathering Validity Evidence by Examining Processes of Responding

Traditionally, validity theory and applied efforts to gather supporting evidence for valid inferences from measures have focused much on quantitative approaches such as correlation between the measured attribute and other attributes. However, it has become obvious that in this approach, an important aspect about the data is missing. As Borsboom, Mellenbergh & van Heerden (2004) specify, validation research should not focus on correlations among attributes but rather, should place attention on the processes that convey the effect of the measured attribute on a test score. In particular, the above authors point to shortcomings of traditional validity approaches, such as the fact that every single concern about psychological testing is relevant and needs addressing and that in social science research, practically everything correlates with everything else. Hence, this approach leaves the researcher without direction or practical guidance. In fact, several authors have pointed out the importance of placing attention on underlying cognitive psychological principles and processes of responding during a task

(Borsboom et al., 2004; Embretson & Gorin, 2001; Kane, 2001; Zumbo, 2005). According to Borsboom et al. (2004), the key aspect of validity involves the causal effects of an attribute on scores. Hence, the locus of evidence for validity resides in the processes that convey this effect. It follows that correlations between scores and other measures can only provide circumstantial evidence. Borsboom et al. (2004) focus on the message that somewhere in the chain of events between item administration and item response, the attribute measured must play a causal role in determining the value of measurement outcomes. Otherwise, the test cannot be valid for measuring the attribute. According to Borsboom et al. (2004), while the view that correlation is a defining feature of validity would mean that everything is valid for everything else to some degree, this problem does not arise in causal theory, since not everything causes everything else. The primary objective in any validation effort should hence be to offer a theoretical explanation of the processes of responding to the items leading up to the outcome.

Zumbo & Zimmerman (1993) called for the replacement of nomological networks with an understanding of the generating process of the variable. They explain that this could include a theory of the response process as such, or could also involve a general theory of the process such as reaction times. This focus would also make for stronger causal models. That is, if we have explanatory models of the source variables and the generating process, then this is the whole of validity; the need to talk about construct validity would dissolve. To obtain an explanatory focus, cognitive theories are the tool of choice (Zumbo, 2005). Zumbo also states his view that ‘validity’ per se is on its own, while the whole is measurement quality.

Kane (2001) argues that validation requires an extended analysis of evidence depending on a clear statement about the intended interpretation. In practical terms, one could ask, for example, whether scales constructed from rating policy documents are justifiably used as quantitative indicators reflecting the strength of a written policy document. According to Kane, a validity argument would involve various kinds of evidence pertaining to the different parts of the interpretive argument. That is, questionable interpretations could be strengthened by improving the measurement procedures as well as objects of measurement. It is here that underlying processes of item responding are potentially useful in illuminating how raters go from the item to producing the observed outcome, a rating score. In this light, qualitative methodologies such as a TA protocol can provide an avenue to the improvement of a) the measurement procedure, b) the objects of measurement and c) the measurement tool.

The Explanation-Centered View of Validity

As Zumbo (2005) emphasizes, when reflecting on validity, one is concerned with quality. In qualitative research in particular, the researcher deals with the 'crisis of legitimacy' (Rapley, 2007). This issue is concerned with questioning the two key positivist notions about what quality traditionally means in research: 'reliability' and 'validity'. These terms traditionally implied that science, including social science, should produce explanation, universal truth through the process of generating objective knowledge. In this perspective, then, validity refers to "nothing less than truth, known through language referring to a stable social reality" (Seale, 1999, p. 34). However, qualitative researchers in particular have warned that this expectation is unattainable, since there is not 'a truth' but rather multiple and possibly contradictory truths or versions thereof. In addition, one must acknowledge that language does not refer to a stable reality but produces multiple possible understandings of what is real (Rapley, 2007).

In a recent overview on the concept of validity, Zumbo (2009) articulates an explanation - oriented view with a focus on context and pragmatics. This explanation-centered validity view is essential for the endeavour of delineating the cognitive processes that underlie responses. That is, if one can understand the variation in an indicator by illuminating the process of how respondents work from the item to the response, this would go a long way in bridging the inferential gap between the indicator and the construct of interest (Zumbo, 2009). One needs to keep these points in mind while striving to accumulate evidence of quality, which would include validity.

Dissertation Objectives

The objectives of this dissertation were to examine the validity of inferences from school tobacco policy ratings through the lens of the substantive aspect of construct validity (Messick, 1994, 1995), and to delineate potential positive and negative consequences of policy rating score interpretation (Messick, 1994, 1995). In order to examine the process of school tobacco policy rating, it was necessary to have an understanding regarding the nature of the tobacco polices currently in place in Canadian schools. Hence, the purpose was to first characterize these polices in terms of their type, length and comprehensiveness.

Messick (1994, 1995) discusses the substantive aspect of construct validity from the perspective of assessment within personality psychology and achievement testing. In this dissertation, the objective was to apply Messick's view to rating. While validity concerns all types of assessment and the substantive aspect of construct validity is relevant in all measurement, an important difference needs to be highlighted with respect to applying Messick's view to rating. The difference between the application of Messick's view to rating as opposed to personality or achievement testing is the context of the task environment. Specifically, this task environment is contextualized to contain objects of rating (different types of policy documents), rating items, and even ratings already produced. Why is it important to take Messick's view on the substantive aspect of construct validity and apply it to a different setting, the rating of policy documentation? In the context of school tobacco policy rating, this approach would help strengthening the validity of inferences drawn from policy indicators by providing a better understanding of the processes by which the scores where generated. Due to the importance of the task environment, it was necessary to first conduct study one of this dissertation (Zeisser, Lovato, Zumbo, Pullmann & Manske, 2009), where the context of the task environment for policy rating is described in detail.

This dissertation addresses several important practical and theoretical problems that applied researchers frequently encounter in the field of tobacco policy and tobacco control. Specifically, the dissertation problem was to provide answers to questions about the cognitive processes with respect to the raters performing the task of rating written school tobacco policies. Why was it important to study raters' cognitive processes? Cognitive processes inform us about the nature of the data by showing us how people produce item responses. This information is not available via traditional psychometric techniques such as correlation or factor analysis. In addition, scores are a function of stimulus conditions and interactions, such as between the task environment, objects rated, items and persons. The study of the cognitive processes involved illuminates precisely these interactions and contexts. As Cizek et al. (2008) and Zumbo (2009) stress, cognitive processes provide researchers access to important information on score generation and meaning and hence, are valuable in supporting validity claims. Further, Messick (1994, 1995) points out that score validation is an empirical evaluation of meaning and consequences of measurement; hence, validity is closely linked to score meaning.

This dissertation draws on the validity work by Messick (1994, 1995), who also discusses the consequential aspect of construct validity. This aspect includes evidence and rationales for evaluating intended and unintended consequences of score interpretation and use. For example, social consequences may be either positive, such as improved educational policies based on comparisons of student performance, or negative when associated with bias in scoring, interpretation or unfairness in test use (Messick, 1995). With respect to adverse consequences, the primary measurement concern is that any negative impact on individuals or groups must not stem from any source of test invalidity, such as construct-irrelevant variance or construct underrepresentation. Therefore, in this dissertation, it was also important to delineate the potential positive and negative consequences of policy rating outcomes in order to be able to arrive at appropriate recommendations for future directions.

In the area of tobacco control and prevention research, there is a strong reliance on written school tobacco control policies for deriving indicators to predict student smoking; researchers desire to use these policy indicators in statistical models. However, while indicators derived from rating written tobacco control polices are desirable to predict smoking outcomes, they also have inherent challenges and limitations. Specifically, in the area of tobacco control, researchers and policy developers often face pragmatic issues such as: What does this policy score mean? How did it come about? Should we feel confident using these scores as predictors of smoking outcomes? To address these issues, it was felt that by far the richest source of information about tobacco policy rating would be to observe how trained expert raters actually generate their responses via a TA protocol. This information is indispensible if one intends to have a thorough understanding of score meaning. That is, an understanding of the process of rater responding is necessary as a basis to back the validity of inferences made from policy rating scores. The practical issues and current needs in the area of tobacco control research described in the next section formed the rationale for the present study.

Researchers construct quantitative indicators from tobacco policy ratings from the Stephens & English (2002) policy rating rubric. These indicators are frequently used in statistical models to predict youth smoking outcomes and make policy decisions affecting Canadian schools in all provinces. There is current researcher interest (University of British Columbia, University of Waterloo) to construct a new self-assessment tool for Canadian schools to assess their tobacco policies, for which the Stephens & English (2002) policy rating rubric is used as the basis. Applied tobacco researchers, school administrators and policy decision makers require information regarding validity evidence of inferences drawn from ratings based on the Stephens & English (2002) policy rating rubric. Hence, there was a need to generate constructive input to potentially revise the policy rating process for future work. Currently and to the best of our knowledge, no such analysis into the validity of inferences made from this type of rating process is available in the literature, particularly not using a cognitive process model approach via the TA method. Study two of this dissertation aims to fill this research gap. To understand the rating act, one needs to identify the processes involved – but this would not be sufficient. One also needs to understand how respondents organize these processes and how simultaneous processes interact to produce the rating score. Therefore, one would be interested in a cognitive model showing the processes involved in tobacco policy rating; then one could accurately describe their organization and interactions (Hayes & Flower, 1983). A process model also serves the purpose of helping researchers understand the components involved in the rating process. This would enable researchers to speak to some critical validity questions in the field of tobacco research by being able to see relevant issues in a way they did not see them before. Moreover, a cognitive process model so developed would provide researchers with explanations of how the policy rating score came about. This understanding would also help inform future expert rater training by alerting to problem areas and utilizing the resource of existing expert rater knowledge.

A closely related dissertation problem tied into the issue of understanding the meaning of tobacco policy scores and revolves around the various features and components of Canadian tobacco control policies currently in place at schools. It was realized that if one desires to create an informative process model of rating tobacco policies, one needs to first understand what the objects of rating - these policies - actually look like. Specifically, it was necessary to obtain detailed descriptions of the tobacco policies in terms of their essential characteristics, such as length, detail and comprehensiveness. The objects being rated are an important part of the interactions that take place during rating; their characteristics form relevant parts of the task environment. Hence, before the process of tobacco policy rating could be studied, it was necessary to lay a foundation by first conducting a detailed investigation and description of tobacco policy characteristics. This problem was addressed in study one of the dissertation, the policy characterization, by developing the elements of the rating task environment.

This dissertation is, to the best of our knowledge, the first effort to gather validity evidence to support inferences from tobacco policy ratings via an examination of the process of rater responding. It is also the first attempt to provide a series of detailed cognitive models describing what cognitive processes are involved in tobacco policy rating, how these processes are organized and how raters deal with obstacles to policy rating. The results are valuable for policy researchers and those wishing to use scores based on ratings. The process of responding is a viable but underused avenue for obtaining evidence of the validity of inferences one can make from these scores. This is true for policy ratings in the area of tobacco control in particular, but also for ratings in general.

The dissertation has some notable implications that speak to the need for continuing examination of rater-generated qualitative data on the process of responding. First, on a broad level, the importance of various task characteristics for generating different ratings and decisions can help inform policy researchers about what their raters may be considering and focusing on when making these decisions. This information can then be utilized in future rater training, where raters could be alerted to a common tendency to pay attention to particular characteristics when rating polices of various types in order to reduce unfair bias. Raters could also be trained to integrate multiple disjunctive pieces of policy information scattered throughout the rating object, and to infer a ‘yes’ or a ‘no’ decision based on indirect or partial information, as frequently observed in the processes of this study. Further, with respect to future rater training, the expertise and extensive training that the raters in this study had clearly helped them to ‘fill in the gaps’ and cope with rating challenges, namely that many policies lacked information they needed to rate, or had only partial information.

Structure of the Dissertation

This dissertation is written in manuscript – based format following the guidelines of the UBC Faculty of Graduate Studies. The present chapter provides background on the evolving view of validity and sets the stage for the investigation of the validity of inferences from rating scores in the context of tobacco control through the lens of the substantive aspect of construct validity. Chapter two is a tobacco policy characterization study that was conducted as a foundation for the subsequent TA study of policy rating (chapter three). That is, chapter two was needed because it was necessary to get a sense of what the objects of rating – Canadian school tobacco policies - are. Specifically, the purpose of the tobacco policy characterization was to learn in detail about essential features and components of Canadian tobacco control policies currently in place in schools, and to develop the task environment for the subsequent TA study of tobacco policy raters. The following research questions were posed in the policy characterization study: What are key characteristics of written tobacco control policies in place at Canadian schools during the 2006/07 school year? How do these policies compare in terms of prevalence of key school tobacco control policy components across provinces? The policy features of interest were the length and comprehensiveness as well as the type of tobacco control policies (e.g., school-developed or board-developed). For this study, written tobacco control policies (N=196) were obtained from schools and school boards across 10 Canadian provinces in the 2006/07 school year that took part in the Youth Smoking Survey. Important differences between policy types with respect to policy strength are presented and implications for future research and policy input are discussed.

Chapter three of this dissertation is a TA study of expert tobacco policy raters. Chapter three builds on the Canadian school tobacco policy characterization study and focuses on the substantive aspect of construct validity (Messick, 1994, 1995). The following research questions were posed in chapter three: What cognitive processes are involved in the rating process, with raters using the Stephens & English rubric? How are the processes organized? How do cognitive processes in policy raters contribute to our understanding of policy score meaning and the measure, the Stephens & English-based coding scheme? What obstacles do raters face during the rating process? What processes do raters deploy to overcome these obstacles and to complete the rating task? The goal of study two was to identify the information upon which tobacco policy raters concentrate during the rating task, and describe how raters make decisions from that information. In addition, the aim was to make inferences about reasoning processes that raters used to resolve difficult rating items. Study two had the objectives to a) identify and describe categories and themes about the cognitive processes of rater responding while coding policies, b) describe how these processes are structured and organized, c) identify and describe obstacles raters encounter during the rating task, and d) identify and describe what raters do to overcome these obstacles.

The TA study aimed to contribute to a novel approach to validity by enhancing the understanding of raters' cognitive processes during their task of coding tobacco policies using a coding rubric based on Stephens & English (2002). The objective was to examine the processes of responding that generate raters' answers. For this purpose, two expert tobacco policy raters were instructed to think aloud during their task of school tobacco policy rating. A series of cognitive models of the response process in policy raters was developed as an approach to gathering evidence for the validity of inferences made from policy rating scores. Implications for tobacco policy development and for gathering validity evidence via an explanation- focused approach are discussed. Chapter three, in addition to providing results from the TA study and developing cognitive models of rater responding, also provides an extensive discussion about validity and validation, with a specific focus on the validity of inferences from ratings. A new conceptualization by Zumbo (2009) of validity as contextualized and pragmatic explanation is discussed in detail, putting this approach into the context of tobacco policy research. In chapter four, the closing chapter, the implications of the finding from this dissertation are discussed and novel contributions are presented.

References

Adams, M. L., Jason, L. A., Pokorny, S., & Hunt, Y. (2009). The relationship between school policies and youth tobacco use. *Journal of School Health, 79*(1), 17-23.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10*, 67-78.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Boyce, J. C., Mueller, N. B., Hogan-Watts, M. & Luke, D. A. (2009). Evaluating the strength of school tobacco policies: the development of a practical rating system. *Journal of School Health, 79*(10), 495-504.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Cronbach, L. J. & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Embretson, S. E. & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

Evans-Whipp, T., Beyers, J. M., Lloyd, S., Lafazia, A. N., Toumbourou, J. W., Arthur, M. W., et al. (2004). A review of school drug policies and their impact on youth substance use. *Health Promotion International, 19*(2), 227-234.

Goldman, B. A. & Mitchell, D. F. (2003). *Directory of unpublished experimental mental measures (8th ed.).* Washington, DC: American Psychological Association.

Hayes, J. R. & Flower, L. S. (1983). Uncovering cognitive processes in writing: An introduction to protocol analysis. In P. Rosenthal, L. Tamor, & S. A. Walmsley, (Eds). *Research on writing: Principles and methods.* New York: Longman.

Hogan, T. P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802-812.

Hubley, A. M. & Zumbo, B. D. (1996). A dialectic on validity: where we have been and where we are going. *Journal of General Psychology, 123*, 207-215.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4, Measurement Update for the 21st Century), 319-342.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lovato, C. Y., Sabiston, C. M., Hadd, V., Nykiforuk, C. I. J. & Campbell, S. H. (2007). The impact of school smoking policies and student perceptions of enforcement on school smoking prevalence and location of smoking. *Health Education Research, 22*(6), 782-793.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (Ed.). (1989). Validity. In R. Linn (Ed.), *Educational Measurement. (3rd Ed.)*, New

York, N. Y.: Macmillan.

Messick, S. (1994). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. (Research Report No. RR-94-45). Educational Testing Services, Princeton, NJ.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). *Standards of validity and validity of standards in performance assessment.* Educational Testing Services, Princeton, NJ.

Messick, S. (1998). Test validity: A matter of consequence. In B. D. Zumbo (Ed.), *Validity theory and the methods used in validation: perspectives from the social and behavioral sciences.* (pp. 35-44). Netherlands: Kluwer Academic Press.

Milio, N. (2001). Glossary: healthy public policy. *Journal of Epidemiology and Community Health, 55,* 622-623.

Moore, L., Roberts, C. & Tudor-Smith, C. (2001). School smoking policies and smoking prevalence among adolescents: multilevel analysis of cross-sectional data from Wales. *Tobacco Control, 10*(2), 117-123.

Murnaghan, D. A., Sihvonen, M., Leatherdale, S. T., & Kekki, P. (2007). The relationship between school-based smoking policies and prevention programs on smoking behavior among grade 12 students in Prince Edward Island: A multilevel analysis. *Preventive Medicine, 44*(4), 317-322.

Nykiforuk, C. (2004). Municipal tobacco bylaws: Use of geographic information systems to explore relationships between local ETS policy and community characteristics. Unpublished dissertation.

Ordonez, A. (2007). Prevalence of bullying among elementary school children as a function of comprehensiveness of anti-bullying policies and programs in the school. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 67*(8-A), pp. 2889.

Rapley, T. (2007.). In U. Flick (Ed.), *Doing conversation, discourse and document analysis. (1st ed.)*, Los Angeles, CA: Sage Publications.

Schmid, T. I., Pratt, M. & Howze, E. (1995). Policy as intervention: environmental and policy approaches to the prevention of cardiovascular disease. *American Journal of Public Health, 85,* 1207-1211.

Seale, C. (1999). *The quality of qualitative research.* Thousand Oaks, CA: Sage Publications.

Stephens, Y. D., & English, G. (2002). A statewide school tobacco policy review: Process, results, and implications. *Journal of School Health, 72*(8), 334-338.

Tompkins, N. O., Dino, G. A., Zedosky, L. K., Harman, M., & Shaler, G. (1999). A collaborative partnership to enhance school-based tobacco control policies in West Virginia. *American Journal of Preventive Medicine, 16*(3, Supplement 1), 29-34.

Zeisser, C., Lovato, C. Y., Zumbo, B.D., Pullman, A. & Manske, S. (2009). A Descriptive and Comparative Analysis of Canadian School Tobacco Control Policies. Poster presented at the National Conference on Tobacco or Health, Montréal QC.

Zumbo, B. D., & Zimmerman, D. W. (1993). Alternatives to classical statistical procedures. *Canadian Psychology, 34*, 381-383.

Zumbo, B. D. (2005.). Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing. (Invited paper, Samuel J. Messick Memorial Lecture Award). Ottawa, Canada.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 45-79). The Netherlands: Elsevier Science B.V.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validity practice. In R. W. Lissitz, (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65-82). IAP- Information Age Publishing. Inc.: Charlotte, NC.

CHAPTER TWO: A DESCRIPTIVE AND COMPARATIVE ANALYSIS OF CANADIAN SCHOOL TOBACCO CONTROL POLICIES

Introduction and Literature Review

Adolescent smoking rates continue to be a major public health concern in Canada. School-based smoking prevention programs and tobacco control policies are widely used in adolescent tobacco control because schools are established environments wherein students' behaviours can be targeted and reinforced (Alexander, Piazza & Mekos, 2001; Lovato, Sabiston, Hadd, Nykiforuk & Campbell, 2007; Adams, Jason, Pokorny & Hunt, 2009). Research examining the effect of school tobacco policies on adolescent smoking behaviour has been mixed (Alexander et al., 2001; Lovato et al., 2007; Adams et al., 2009; Moore, Roberts, Tudor-Smith, 2001) but suggests that school tobacco policies can have a moderate influence on reducing smoking (Alexander et al., 2001; Lovato et al., 2007; Adams et al., 2009; Lipperman-Kreda, Paschall & Grube, 2009; Murnaghan, Sihvonen, Leatherdale & Kekki, 2007; Pentz, Brannon, Charlin, Barrett, MacKinnon & Flay, 1989; Wakefield, Chaloupka, Kaufman, Orleans, Barker & Ruel, 2000; Moore et al., 2001).

[1] A version of this chapter will be submitted for publication. Zeisser, C., Zumbo, B. D., Lovato, C. L., Pullman, A. and Manske, S. A descriptive and comparative analysis of Canadian school tobacco control policies.

Effective components of school tobacco policies have previously been examined in the literature. Some evidence suggests that effective tobacco control policies are those that are comprehensive and cover a wide range of areas, including awareness on dangers of tobacco use, smoke-free areas, prohibition, education, and cessation (Pentz et al., 1989; Wakefield et al., 2000; Nykiforuk, Campbell, Cameron, Brown & Eyles, 2007). There is also evidence indicating that more comprehensive and strictly enforced tobacco control policies are associated with reduced smoking (Adams et al., 2009; Pentz, et al., 1989; Wakefield et al., 2000; Nykiforuk et al., 2007). Other components recognized in the literature include smoking prevention and cessation access which appear to be more effective than punishment in reducing youth smoking rates (Pentz et al., 1989).

Several studies have used rating systems to evaluate tobacco control polices. Researchers have rated municipal smoking bylaws for the degree of comprehensiveness, restrictiveness and enforcement provisions (Nykiforuk et al., 2007; Nykiforuk, 2004) to understand policy outcomes (Tworek, Sandoval, Thompson, Harper, Slater & Chaloupka, 2006) and to evaluate bylaw restrictiveness (Stephens , Pederson, Koval & Macnab, 2001). Other researchers have developed a rating tool to evaluate workplace smoking policies by identifying the presence or absence of policy components (Glasgow, Boles, Lichtenstein & Strycker, 1996). Tompkins, Dino, Zedosky, Harman & Shaler (1999) conducted a tobacco control policy characterization of 55 county tobacco policies in West Virginia, using a coding protocol to enable judgment about whether a policy addressed particular elements specified by the CDC Tobacco Guidelines. In the school context, Pentz et al. (1989) characterized school tobacco control policies in California according to the degree to which they restricted student smoking on school grounds, and whether the school had a smoking prevention plan. The policies were also rated for comprehensiveness according to the number of components addressed in the policy (e.g. prevention emphasis and punishment emphasis). In New York State, Stephens & English (2002) developed a coding rubric to provide policy reviewers a tool to assess degrees of difference among tobacco control policies in terms of their strength and to systematically quantify these differences. Furthermore, school tobacco policy scores have previously been used in statistical models to evaluate the impact of policies on student smoking rates (Lovato et al., 2007; Murnaghan, Sihvonen, Leatherdale & Kekki, 2007). Such indicators are desired by policy makers at the school-and board level to design effective tobacco control polices. Despite this previous work evaluating tobacco policies, there remains a lack of research that describes the status and characteristics (e.g., length and comprehensiveness) of Canadian school tobacco control policies.

The purpose of the current study was to describe the status and characteristics of school tobacco control policies in Canada and compare the prevalence of key policy components across provinces. A related purpose was to inform the design of a subsequent think-aloud study of tobacco policy rating by delineating school tobacco policy characteristics as the input into the task environment of rating policies. The results of this study will inform policy-makers about the status of school tobacco policies in Canadian elementary and secondary schools. It will also help researchers and policy makers develop tobacco policy monitoring and assessment tools.

Methods

Data sources

The current study collected written school tobacco policies in tandem with the Health Canada sponsored Youth Smoking Survey (YSS). Schools with students in grades 5 to 12 and located in the 10 Canadian provinces (British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, Newfoundland and Labrador, New Brunswick, Nova Scotia and Prince Edward Island) were eligible to participate. Schools in the Yukon, Nunavut and the Northwest Territories, as well as schools on First Nation Reserves, special schools and schools on military bases were excluded from the sample. A stratified, multistage sample design was used. Within each province schools were stratified by health region smoking rate (above or below median) and by school type (elementary or secondary). In each province, schools were randomly selected to participate with probabilities proportional to the total enrollment in their school boards. Private schools were sampled proportional to the number of students enrolled in private schools in the province compared to the total in public schools. School administrators at each participating school were asked to provide all written documentation pertaining to smoking policies at their school that were in effect during the 2006/07 school year.

Procedure

The first step was to describe characteristics of Canadian school tobacco control policies. The descriptive characterization started with an overall assessment of prevailing characteristics of the collected school policies. Polices were read and key components were selected based on elements of a comprehensive tobacco control policy as stated by the World Health Organization (WHO, 1999), the Centre for Disease Control's (CDC) Tobacco Guidelines (1994), and Stephens & English (2002), see coding scheme presented in Table 1. Each school's written tobacco policy was read and coded by a trained researcher (the first author) according to the criteria in Table 1. A numeric value according to these criteria was assigned to each policy component. A second trained coder read and coded a subset of policies (n=10) in order to provide feedback and another perspective on the policy characterization. Initial inter-rater agreement was 85%. Any discrepancies were discussed until consensus was reached and the policy characterization criteria were finalized.

Table 1. Variables and Coding Used In Characterizing Policies

Variable	*Coding*
Policy Characteristics	
Policy type	1 = school-developed (e.g., stand-alone policy document, student handbook, parent handbook, code of conduct, newsletter) 2 = board-developed
Policy length	1 = short (one paragraph or less) 2 = long (several paragraphs)
Policy comprehensiveness	1 = minimal detail - only one policy component, e.g., only the "No Smoking" statement 2 = medium detail, two components, e.g., policy mentions rules and consequences of violation 3 = extensive detail, three or more components, e.g., rules, smoke-free spaces, consequences of violation, to whom the school policy applies, prevention education and cessation help
Key Policy Components	
Prohibition (existence of a clear set of tobacco rules - 'zero tolerance')	0=absent, 1=present
Enforcement (specified actions taken when students violate)	
Parents (involving parents' input in smoking prohibition, prevention, enforcement or cessation efforts)	
Smoking prevention education (school has prevention education programs, curriculum)	
Cessation access (school offers smoking cessation programs)	
Location of prohibition (buildings, property, vehicles)	

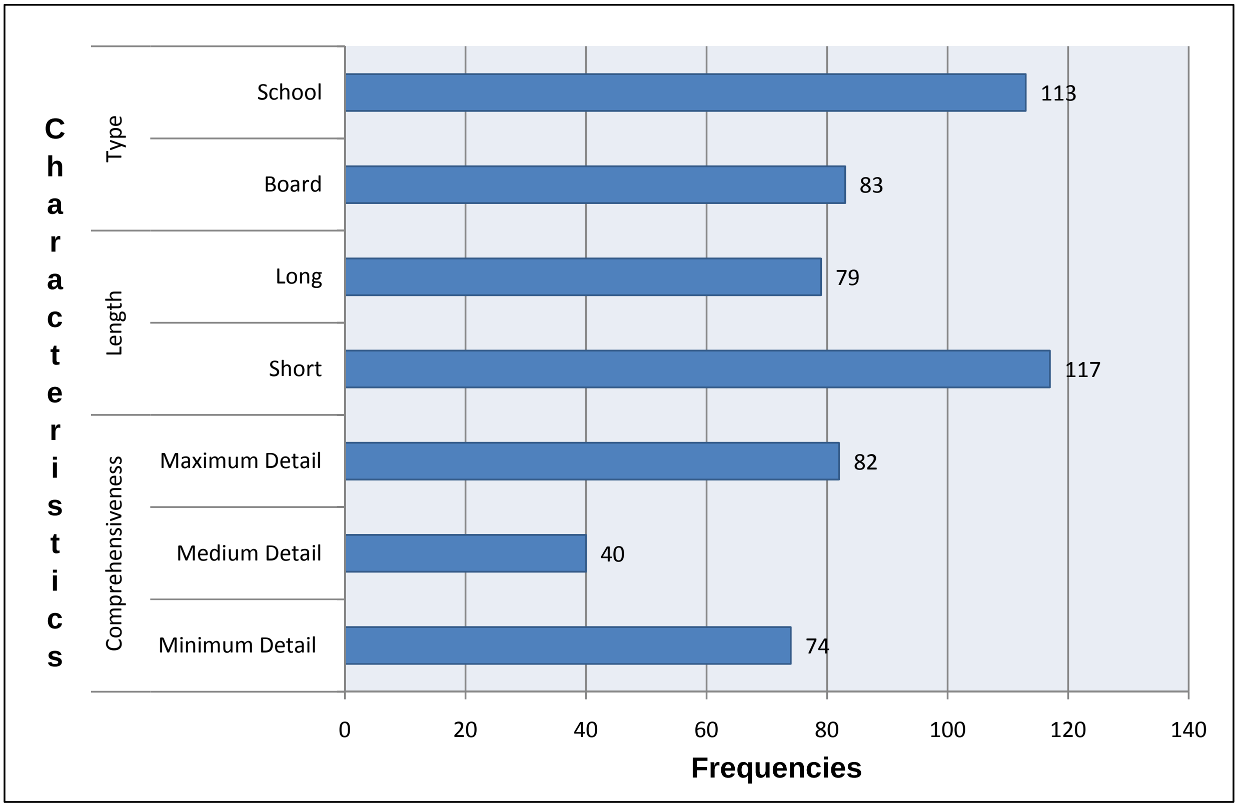
Analysis

Univariate distributions of policy characteristics (length, comprehensiveness, type) were described to assess their range and variability. Proportions were obtained for presence or absence of policy components with binary scoring (0=component not addressed, 1=component addressed). The focus here was on depicting what school tobacco control policies looked like in Canada, according to the specified components of a strong tobacco control policy. Descriptive analyses were also performed separately to discern differences between school-developed and board-developed polices with respect to policy components. A summary analysis was conducted to show how many policies addressed each of the key policy components.

Results

Of the 313 schools recruited as part of the Youth Smoking Survey, 196 schools provided policies (63% response rate). The majority of tobacco control policies were school-developed policies as compared to board-developed policies (Figure 1). Surprisingly, only 19 polices were categorized as the school's own policy - a stand-alone document. The remaining school-developed policies were excerpts from student-or parent handbooks (n=58), codes of conduct (n=17) or other documents (n=19). Figure 1 shows the distributions of policy type, length and comprehensiveness.

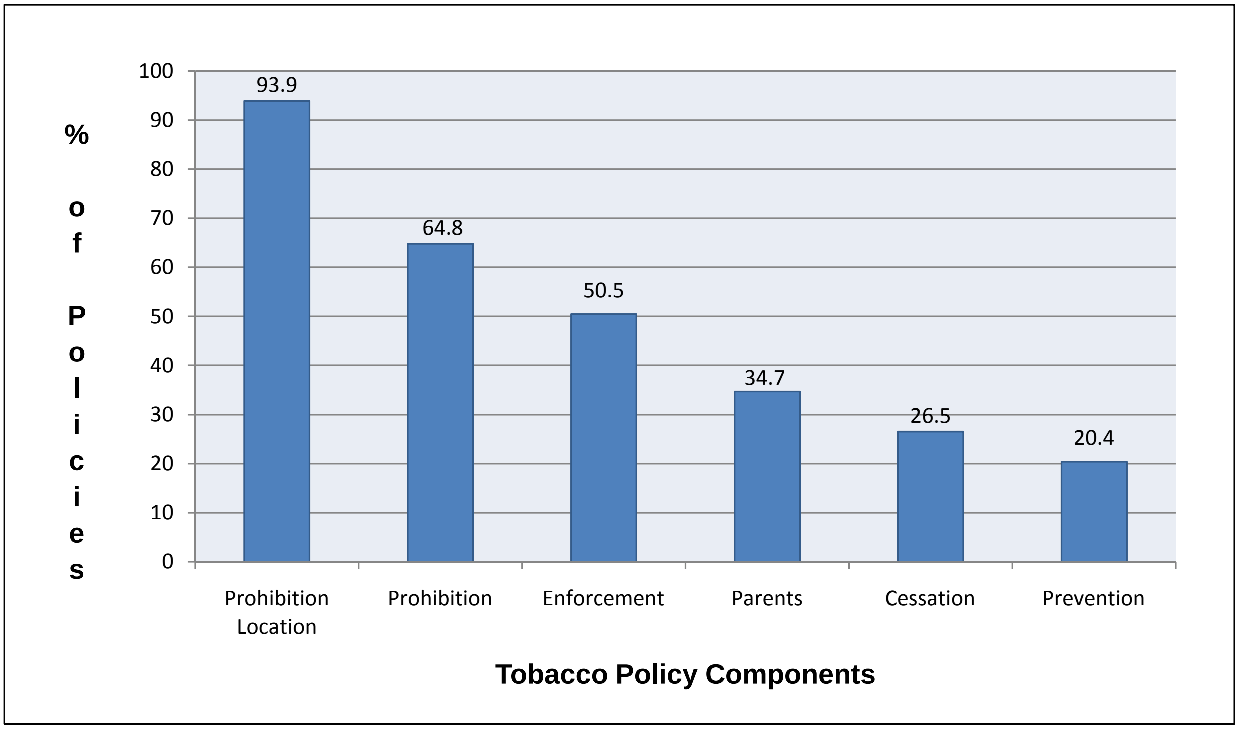
Figure 1. Distribution of tobacco control policy characteristics.



Summary of tobacco control policy components

A summary analysis across all policies revealed that key policy components were addressed to varying degrees. Figure 2 shows the percentage of tobacco control policies that addressed key policy components investigated in the present study.

Figure 2. Summary analysis: percentage of tobacco control policies addressing key policy components.



Differences in types of policy, different policy foci

Policy components varied with respect to policy type (e.g., school-developed versus board-developed). Very few school-developed policies had a prevention education (3%, n=4) or smoking cessation component (11%, n=13). In comparison, board-developed policies addressed prevention education and smoking cessation access more frequently, (43%, n=36 and 47%, n=39 respectively). In contrast, school-developed policies (37%, n=42) addressed a parent involvement component more often than board developed policies (31%, n=26). A statistical comparison showed that the differences between school-developed and board-developed policies were significant with respect to addressing prevention education (χ^2 (1, (N=196) = 46.74, p = .001) and smoking cessation access (χ^2 (1, (N=196) = 30.91, p = .001). Policy type with respect to addressing a parent involvement component was not significant.

Discussion

This study examined school tobacco control polices from 10 Canadian provinces by describing characteristics and key components of policies currently in place. Results from the current study indicate that these policies vary greatly in strength.

A compelling finding from this study was that the smoking prevention education and smoking cessation access elements were absent from the majority of school-developed policies currently in place in Canadian schools despite the fact that they have been identified as a key policy component (Pentz et al., 1989; CDC, 1994). Furthermore, board-developed policies were stronger than school-developed policies with respect to smoking prevention education and smoking cessation access. This result emphasizes the need for schools to focus more strongly on programs that support and reinforce tobacco control policies.

Our results are in line with current knowledge related to tobacco control indicating a need for much stronger emphasis on preventive actions reinforcing a smoke-free environment, e.g., through tobacco prevention education (Murnaghan, Leatherdale, Sihvonen & Kekki, 2008; Pentz et al., 1989; CDC, 1994). Our results are also consistent with the findings by Tompkins et al. (1999), who reported that only 11% of West Virginia county tobacco policies addressed the need for tobacco use prevention education. In current school-developed policies there seems to be more emphasis on punitive actions and consequences after violations have occurred. This is of interest because smoking prevention education and cessation access have been recognized in the literature as more effective than punishment in reducing youth smoking rates (Pentz, et al., 1989). Hence, these elements need to be addressed more strongly in Canadian school tobacco control policies. This is particularly the case for policies developed at the school level versus the board level.

Further findings regarding tobacco policy characteristics are comparable to those of Tompkins et al. (1999), in that board-developed policies (or county policies, in the West Virginia study) in general were more comprehensive than policies developed at the school level (e.g., addressing more key policy components). In the present study, board-developed policies also tended to be longer than school–developed tobacco control policies.

The present study had several limitations. First, the results speak only to the 10 Canadian provinces. The Canadian territories (Yukon, Nunavut and Northwest Territories) were not represented in this study. Further, the nature of the present study was predominantly descriptive and exploratory and hence, no inferences can be drawn with respect to policy effectiveness. It is important to note that several schools did not provide a written policy (63% response rate) because they did not have a policy or because we were unable to collect the policy after repeated attempts. In addition, the range of policy components identified and described here is not exhaustive; other relevant elements could be determined and examined. Future research should continue to monitor the school policy environment, particularly questions related to the implementation of policies.

The results from this study provide information regarding the status of school-based tobacco control policies. In addition, results may be useful in guiding schools to update or strengthen their school tobacco control policies and focus on elements that need improvement, specifically the policy components of tobacco prevention education and smoking cessation access. Finally, the findings provide new information to tobacco researchers interested in instruments that can be used by schools for self-assessment purposes.

The findings of the present study indicate that school tobacco control policies in Canada vary greatly in length and comprehensiveness with very few policies addressing smoking prevention or cessation programming. Schools are encouraged to develop comprehensive policies that not only address prohibition but also enforcement, parental involvement and prevention and cessation programming.

References

Adams, M. L., Jason, L. A., Pokorny, S., & Hunt, Y. (2009). The relationship between school policies and youth tobacco use. *Journal of School Health, 79*(1), 17-23.

Alexander, C., Piazza, M. & Mekos, D. (2001). Peers, schools and adolescent cigarette smoking. *Journal of Adolescent Health, 29*(1), 22-30.

Barnett, T.A., Gauvin, L., Lambert, M., O'Loughlin ,J., Paradis, G. & McGrath, J. J. (2007) The influence of school smoking policies on student tobacco use. *Archives of Pediatrics & Adolescent Medicine. 16*1(9):842-848.

Bonnie, R. J., & Lynch, B. S. (1994). Time to up the ante in the war on smoking. *Issues in Science & Technology, 11*(1), 33-37.

Brownson, R. C., Koffman, D. M., Novotny, T. E., Hughes, R. G., & Eriksen, M. P. (1995). Environmental and policy interventions to control tobacco use and prevent cardiovascular disease. *Health Education Quarterly, 22*(4), 478-498.

Centers for Disease Control and Prevention (1994). Guidelines for School Health Programs to Prevent Tobacco Use and Addiction. *MMWR Morb Mortal Wkly Rep. 43*(RR-2).

Darling, H., Reeder, A. I., Williams, S. & Mcgee R. (2006). Is there a relation between school smoking policies and youth cigarette smoking knowledge and behaviors? *Health Education Research; 21*(1), 108-115.

Glasgow, R., Boles, S., Lichtenstein, E., & Strycker, L. (1996). Tobacco policy rating form: A tool for evaluating worksite and tribal smoking control policies. *Tobacco Control, 5*(4), 286-291.

Griesbach, D., Inchley, J. & Currie, C. (2002). More than words? The status and impact of smoking policies in Scottish schools. *Health Promotion International, 17*(1), 31-41.

Lipperman-Kreda, S., Paschall, M. J. & Grube, J.W. (2009). Perceived enforcement of school tobacco policy and adolescents' cigarette smoking. *Preventive Medicine, 48*(6); 562-566.

Lovato, C. Y., Sabiston, C. M., Hadd, V., Nykiforuk, C. I. J. & Campbell, S. H. (2007). The impact of school smoking policies and student perceptions of enforcement on school smoking prevalence and location of smoking. *Health Education Research, 22*(6), 782-793.

Moore, L., Roberts, C. & Tudor-Smith, C. (2001). School smoking policies and smoking prevalence among adolescents: multilevel analysis of cross-sectional data from Wales. *Tobacco Control, 10*(2), 117-123.

Murnaghan, D. A., Sihvonen, M., Leatherdale, S. T., & Kekki, P. (2007). The relationship between school-based smoking policies and prevention programs on smoking behavior among grade 12 students in Prince Edward Island: A multilevel analysis. *Preventive Medicine, 44*(4), 317-322.

Murnaghan, D. A., Leatherdale, S. T., Sihvonen, M., & Kekki, P. (2008). A multilevel analysis examining the association between school-based smoking policies, prevention programs and youth smoking behavior: evaluating a provincial tobacco control strategy. *Health Education Research; 23*(6), 1016-1028.

Nykiforuk, C. (2004). Municipal tobacco bylaws: Use of geographic information systems to explore relationships between local ETS policy and community characteristics. Unpublished dissertation.

Nykiforuk, C., Campbell, S., Cameron, R., Brown, S. & Eyles, J. (2007). Relationships between community characteristics and municipal smoke-free bylaw status and strength. *Health Policy, 80*(2); 358-368.

Pentz, M. A., Brannon, B. R., Charlin, V. L., Barrett, E. J., MacKinnon, D. P., & Flay, B. R. (1989). The power of policy: The relationship of smoking policy to adolescent smoking. *American Journal of Public Health, 79*(7), 857-862.

Poulin, C.C. (2007). School smoking bans: do they help/do they harm? *Drug Alcohol Rev 26*(6), 615-624.

Stephens, T., Pederson, L. L., Koval, J. J., & Macnab, J. (2001). Comprehensive tobacco control policies and the smoking behaviour of Canadian adults. *Tobacco Control, 10*(4), 317-322.

Stephens, Y. D., & English, G. (2002). A statewide school tobacco policy review: Process, results, and implications. *Journal of School Health, 72*(8), 334-338.

Tompkins, N. O., Dino, G. A., Zedosky, L. K., Harman, M., & Shaler, G. (1999). A collaborative partnership to enhance school-based tobacco control policies in West Virginia. *American Journal of Preventive Medicine, 16*(3, Supplement 1), 29-34.

Tworek, C., Sandoval, A., Thompson, M., Harper, D., Slater, S. & Chaloupka, F. (2006). SmokeLess States: Evaluating Tobacco Legislation to Understand Policy Outcomes. Presented at the American Public Health Association, Boston, MA.

Wakefield, M.A., Chaloupka, F.J. & Kaufman, N.J., Orleans, C.T, Barker, D.C. & Ruel, E.E. (2000). Effect of restrictions on smoking at home, at school, and in public places on teenage smoking: Cross sectional study. *British Medical Journal, 321*, 333-337.

World Health Organization (WHO) Western Pacific Region (1999). Tobacco Control Initiative. Retrieved June 2 2009 from http://www.wpro.who.int/sites/tfi/.

CHAPTER THREE: AN INVESTIGATION INTO SCHOOL TOBACCO POLICY RATING FROM THE PERSPECTIVE OF THE SUBSTANTIVE ASPECT OF CONSTRUCT VALIDITY

Introduction and Literature Review

Studies in policy research routinely aim to quantify policies and apply rating systems to do so. These rating or coding systems are essentially tools for characterizing policies. The task of policy quantification via coding or rating schemes has become an increasingly used empirical approach in various research contexts such as clean-air bylaw coding (Nykiforuk, 2004), school drug policy (Evans-Whipp, Beyers, Lloyd, Lafazia, Toumbourou, Arthur et al., 2004) and anti-bullying policies in schools (Ordonez, 2007). Policy ratings are also frequently conducted in the context of school tobacco control (e.g., Lovato, Sabiston, Hadd, Nykiforuk & Campbell, 2007; Murnaghan, Sihvonen, Leatherdale & Kekki, 2007; Adams, Jason, Pokorny & Hunt, 2009; Boyce, Mueller, Hogan-Watts & Luke, 2009; Tompkins, Dino, Zedowsky, Harman & Shaler, 1999; Moore, Roberts & Tudor-Smith, 2001).

[2] A version of this chapter will be submitted for publication. Zeisser, C., Zumbo, B. D., Lovato, C. Y. and Young, R. An investigation into school tobacco policy rating from the perspective of the substantive aspect of construct validity.

Researchers use rating schemes to quantify variables of interest, such as the strength of a policy, because they are interested in exploring the connection between the policy and outcomes such as youth smoking behaviours. Tobacco prevention researchers often desire to use policy scores as explanatory variables; from a statistical point, therefore, it helps to characterize policies and then relate them to outcomes of interest. Hence, characterizing and quantifying policies are important activities for today's researchers.

In the field of tobacco use prevention research, a commonly used rating system is the Stephens & English (2002) rubric, intended to measure school tobacco policy strength. The rubric is used in published research (e.g., Lovato et al, 2007; Adams et al., 2009; Boyce et al., 2009). Using existing school tobacco control polices, Stephens & English developed a coding rubric to provide policy reviewers with a tool to assess degrees of difference among tobacco control policies in terms of their strength and to systematically quantify these differences. The rubric covers several areas relevant to school tobacco policies (e.g., policy development, enforcement, prevention and cessation). Raters assign codes for presence (1=yes) or absence (0=no) of specific policy criteria to arrive at a total score to reflect the strength of a school tobacco policy. The rubric identifies characteristics that must be included for a policy to be considered strong. For example, a school tobacco control policy is considered to be strong if it was developed with students, is comprehensive, consistently enforced and addresses smoking prevention and cessation (Stephens & English, 2002; Lovato et al., 2007). The Stephens & English (2002) coding rubric contains the following five school tobacco policy components:

1. Developing, Overseeing and Communicating the Policy (15 possible points)
2. Purpose and Goals of the Policy (9 possible points)
3. Tobacco-Free Environments (18 possible points)

4. Tobacco Use Prevention Education (27 possible points)
5. Assistance to Overcome Tobacco Addiction (12 possible points)

According to Stephens & English, these components need to be present in order for a school tobacco policy to be considered strong. Each of these components is assessed with specific coding criteria (items). These five policy components are used to create subscales with scores for each component as well as a total policy score over all components. The maximum possible final score is 81 points (Stephens & English, 2002). Small modifications were made to the Stephens & English coding rubric for use by the Tobacco Research Unit (TRU) at the University of British Columbia. These modifications were made to reflect a stronger emphasis on tobacco-free environments and included the addition of the subscales prohibition, strength of enforcement and characteristics of enforcement.

Quantification of policy has the advantage of enabling decision makers and researchers to identify and focus on important practical issues that need to be addressed. However, there are two main challenges in quantification via ratings: the availability and choice of a coding instrument and the ability to draw valid inferences from ratings by a clear understanding of what the policy rating score means. In this paper I focus on the second problem through the lens of the substantive aspect of construct validity (Messick, 1994, 1995). The purpose of this study is to investigate the validity of inferences from school tobacco policy ratings based on the Stephens & English (2002) rubric by focusing on cognitive processes in raters via the think-aloud (TA) method. The paper has two main foci: i) study processes of rater responding as a source of validity evidence via a TA protocol, and ii) examine these processes of responding in the specific context of Canadian school tobacco policy ratings.

This paper begins with a presentation of the need to investigate cognitive processes as sources for validity evidence, followed by an introduction of the use of thought processes and their significance in validity studies. This will set the stage for a detailed discussion of respondents' thought processes to gather evidence for the validity of inferences from scores. In particular, I will focus on the benefits of utilizing respondents' cognitive processes to support validity claims above and beyond traditionally used psychometric approaches. The focus will then shift toward issues regarding the validity of inferences from rating data. The rationale for this study is presented before the background of how the need for this particular approach to validity research arose within the context of Canadian school tobacco policy rating. Next, I describe the TA method in detail, along with some background on its traditional use and how it can be applied in the area of policy rating with the aim to obtain information pertinent to the validity of inferences from ratings. I also describe potential limitations of the TA method, keeping in mind the methodological significance of the TA method and how its focus on respondents thought processes can help researchers better understand the rating process. I will describe protocol analysis (Ericsson, 2002) and the use of the TA method specifically for rating tasks, followed by the results and descriptive information. Obstacles to rating are described in detail using extensive examples of excerpts from the transcripts to demonstrate how raters approached rating obstacles and applied coping strategies in order to complete the rating task. This information is integrated into a series of cognitive process models. I present conclusions and implications for future research, along with the limitations of the present study. In the following, I briefly review the changes in researchers' approach to gathering validity evidence.

Changing Approaches to Gathering Validity Evidence

Traditionally, researchers have examined scores and measures using standard psychometrics, such as calculating correlations and validity coefficients. This is because historically, the primary focus in construct validation has been placed on internal and external test structures in form of an appraisal of theoretically excepted relationships among scores or between scores and other measures. However, researchers such as Messick (1995), Kane (2001, 2006) and Zumbo (2009) recently have explored and proposed an examination of the cognitive processes that generate respondents' answers as sources for validity evidence. Messick (1994) also suggests that direct probes and modeling of processes underlying responses - accessible via the TA approach - are more illuminating of score meaning than correlation coefficients with other measures. The following section describes Messick's view on validity, with a particular focus on the substantive aspect of construct validity.

The Substantive Aspect of Construct Validity

Messick (1994, 1995) discusses six aspects of construct validity: content, substantive, structural, generalizability, external and consequential. The content aspect of construct validity includes evidence of content relevance, representativeness and technical quality. The structural aspect appraises fidelity of the scoring structure to the structure of the construct domain of interest. The generalizability aspect examines the degree to which score interpretations generalize across populations, tasks and settings. The external aspect includes convergent and discriminant evidence from multitrait-multimethod studies as well as evidence of criterion relevance and applied utility. The consequential aspect appraises value implications of score interpretation as a basis for action and the actual and potential consequences of test use in regard to fairness and bias. In the substantive aspect of construct validity, Messick (1994, 1995) focuses our attention towards theoretical rationales for the observed consistencies in test responses. Here, Messick suggests the use of theory and process models as important, albeit rarely explored avenues for gathering evidence in support of construct validity. The substantive aspect focuses on the role of substantive theories and process modeling in identifying domain processes to be revealed in assessment tasks (Messick, 1989, 1994 and 1995). Hence, this aspect emphasizes the need for empirical evidence of response consistencies reflective of domain processes. This evidence could be obtained via a TA protocol during task performance.

In the substantive aspect of construct validity, two important points are involved: One is the need for tasks that appropriately sample domain processes; the other is the need to move beyond traditional professional judgment of content to gather empirical evidence that the sampled processes are indeed engaged in by respondents during the assessment task. Hence, the substantive aspect adds the need for the process representation of the construct of interest and the degree to which these processes are reflected in construct measurement (Messick, 1995). The following section reviews the reporting practices with respect to validity evidence and focuses in particular on the reporting of cognitive processes as a source of validity evidence.

The Status of Reporting on Cognitive Processes as Sources of Validity Evidence

Cizek, Rosenberg and Koons (2008) note the scarcity with which cognitive processes are examined in the context of gathering validity evidence and encourage using more process models. Cizek et al. (2008) and Hogan & Agnello (2004) point out that some sources of validity evidence are essentially ignored in the majority of published reports. Specifically, Cizek et al. (2008) note that the majority of articles reports only on construct validity (58.0%), criterion-related/concurrent validity (50.9%) and some on content validity (48.4%). Only 1.8% of articles reported on response processes as type of validity evidence. However, as Cizek et al. (2008) and

Zumbo (2009) emphasize, the study of respondents' cognitive processes when they answer items is an invaluable tool for researchers to access crucial information to help them support their validity claims. Further, the Standards for Educational and Psychological Testing (APA, AERA and NCME, 1999) call for the use of process models in validation practice. In short, while the importance of investigating cognitive processes of responding is often talked about, it is seldom done. Information gathered from such investigations can supply valuable insights to add to the validity argument. Hence, one needs to illuminate the process of responding in general, and the process of rating specifically in the context of the present research. The present study does so with a focus on the process of rating Canadian school tobacco policies. The approach used to accomplish this goal is the TA method. In the next section, I highlight the value of studying response processes above and beyond traditional psychometrics to gather validity evidence.

Importance of Studying Response Processes Above and Beyond Traditional Psychometrics

In health - and educational studies, researchers require evidence about the validity of inferences from scores based on the measurement tools they are applying. Validity evidence is often gathered through quantitative methods such as correlating measures with outcomes or accepted 'gold standards'. However, gathering validity evidence on solely quantitative grounds is not sufficient, since this approach does not inform the researcher about the actual process of responding that underlies participants' answers to items or tasks. Specifically, if one is interested in making evidentiary claims about the validity of scores from educational and other assessments, one needs to understand examinees' thinking processes. How does knowledge of respondents' cognitive processes tie in with validity evidence supporting score inferences? The cognitive process of responding is important for the validity of inferences one wishes to draw, since these processes inform us about the nature of the data by showing how responses to items are being produced (Messick, 1994, 1995; Embretson & Gorin, 2001). As Messick (1989 and 1995) emphasizes, validity is an overall evaluative judgment of the extent to which theory and empirical evidence support one's interpretations of and actions based on scores. Score validation is an empirical evaluation of the meaning and consequences of measurement (Messick, 1994). As such, validity is closely linked to the meaning of scores. Further, scores are a function of the stimulus conditions and their interactions (e.g., task environment, objects of measurement, items and persons responding).

Messick (1994) further states that a fundamental aspect of construct validity is construct representation; the goal here is to identify - through analyses of cognitive processes - the theoretical mechanisms underlying task performance. This is done primarily by decomposing the task into component processes. These processes can then be assembled into functional models or process theory (Embretson, 1983). Construct representation refers to the relative dependence of responses during a task on the processes, strategies and knowledge of respondents (including meta-cognitive or self-knowledge) that are implicated in task performance. It follows that cognitive process models are valuable tools for examining these relations since they help shed light on the nature of the scores.

Another aspect of cognitive processes closely relates to validity of score inferences via better explanation. It is these cognitive processes that one needs to understand if one aims to marshal explanations for observed variations in scores. Understanding and explanation are two central points in a newly articulated broader view of validity by Zumbo (2009). Zumbo argues that in order to be able to make valid inferences from scores, one needs to understand those processes of responding that cause variation in score outcomes; understanding the processes that generated score variation forms the basis for being able to provide explanations for the scores.

Specifically, understanding why an individual responded a certain way to an item or scored a particular value on a scale would go a long way toward bridging the inferential gap between test scores and constructs (Zumbo, 2009). In this sense, knowledge elicitation from experts by asking them to think aloud during a task is a preferred way of attaining insights into how scores were generated since it provides unique information about the nature of the expert's response. This approach is suitable if one considers Messick's (1994) point that construct validity comprises evidence and rationales supporting the trustworthiness of score interpretation with respect to explanatory concepts that would help account for score variation. As Messick (1995) states, the principles of validity apply to all assessments, whether questionnaire-based, observation-based or rating-based. Hence, the value of validity becomes also obvious in the study of thought processes that raters engage in during their task. The process of validation combining scientific inquiry with rational judgment to justify (or challenge) score interpretation and use becomes relevant in the context of studying thought processes. In addition to insights into the process of responding, Zumbo (2009) also emphasizes that validity is contextualized and pragmatic explanation of variation in one's observations. Hence, explanation-oriented studies aimed at understanding cognitive processes, such as TA studies, need to take into consideration the context of the process observation comprising the task environment.

As Backlund, Skaner, Montgomery, Bring & Strender (2003) emphasize in the health context, verbal reports from TAs provide information about the cognitive process that is not captured by the ratings alone. Examples are how decision rules are applied, what medical knowledge is used in the decision and how different types of information about the patient are attended to and evaluated. Answers to similar questions in the context of rating in general are important, since they would potentially help improve decisions through better teaching and guidelines. The following section will briefly discuss the literature on evaluating the validity of inferences specifically from rating data.

Validity of Inferences from Rating Data

While the use of trained raters has a long-standing history in educational research (e.g., Carpenter, Fennema, Peterson, Chiang & Loef, 1989; Coffman, 1971; Gay & Gallagher, 1976; Moss, Cole & Khampalikit, 1982), surprisingly, the validity of inferences from rating data has received relatively little attention (Harwell, 1999). The use of trained raters is not without problems. The rating literature frequently points out that ratings are fallible, subject to rater errors or rater effects. For example, Saal, Downey & Lahey (1980) warn that rating fallibilities can occur in four types: a) rater severity, b) halo, c) central tendency and d) restricted variability. Rater severity is present when a rater consistently rates too harshly or too leniently (Coffman, 1971). Halo rating effects occur when a rater is asked to rate only a part of, for example, an essay but instead rates the essay in a holistic fashion. Central tendency and restricted variability rater effects occur when raters spuriously tend to rate too similarly.

It is well documented that rater effects can affect the reliability of ratings; but they can also affect the validity of inferences made from the ratings. This issue has received too little attention in the literature. Valid ratings are ratings that are accurate reflections and characterizations of whatever is being rated. Since ratings frequently play an important role in high-stakes decisions in various areas, invalid ratings can have serious consequences and should be of great concern (Harwell, 1999). While cognitive processes during rating tasks have been examined via the TA method by other researchers as discussed earlier, there appears to be no study that provides a model of the cognitive processes that produce the ratings. In addition, to the best of our current knowledge, no studies in the area of policy rating or tobacco research seem to have produced such models of responding. The following section clarifies the need for describing the rating process using cognitive process models to address issues regarding our understanding of scores produced by ratings, and the validity of claims one can make from such scores.

The attention validity issues receive in studies involving ratings is variable. For example, in large-scale performance assessments such as the Georgia writing assessment, one strategy is to compare ratings by a committee to benchmark samples. Substantial discrepancies between the committee ratings of the benchmark papers and those assigned by raters raised concerns about validity (Harwell, 1999). In a different study, Baxter, Shavelson, Goldman & Pine (1992) carefully documented their attempts to ensure that performance assessments were representative of the desired domain of skills, showing that ratings distinguished between key groups; they also offered correlations among various measures as validity evidence.

Other types of rating studies focus on validating the performances to be rated or the scoring system used by raters, but pay little attention to the validity of inferences from the ratings themselves (e.g., Bennett, Rock & Wang, 1991; Cannella, 1992; Breland, Danos, Kahn, Kubota & Bonner, 1994; Fuchs, Fuchs, Bentz, Phillips & Hamlet, 1994). While evidence of the validity of inferences one can draw from the scoring system is important, it does not automatically establish the validity of inferences from the ratings. That is, one could have valid representations of the skill domain tested, but inferences from the assigned ratings could be limited due to lack of rater training.

Another group of rating studies does not provide validity evidence of any kind, but simply reports coefficients of inter-rater reliability or agreement (e.g., King, 1992). Harwell (1999) provides two frameworks for gathering evidence for validity of inferences from rating data. One focuses on raters as data collection instruments who should be subject to traditional guidelines for establishing validity evidence, treating raters as interchangeable or as data collection instruments representing parallel forms by virtue of having been calibrated (trained). Another framework employs the notion of internal validity linked to experimental designs to provide such evidence, where raters represent the independent variable and ratings the dependent variable; research has focused on interactions of raters and the characteristics of what is being rated (e.g., Huot, 1990). For rating studies, these experimental design conditions are crucial for establishing internal validity (Messick, 1989).

From this review of research on the evaluation of validity evidence in rating studies, it becomes clear that this issue is not focused on sufficiently; historically, the emphasis in rating studies has been on the magnitude of reliability coefficients, rather than on the validity of inferences from the ratings. Only few studies employ the TA method; to the best of our knowledge, none appears to focus on cognitive processes during rating. While cognitive process models have been developed and applied in the areas of reading comprehension (e.g. Anderson & Pearson, 1984; Baker & Brown, 1984; Beach & Hynds, 1991; DeBeaugrande, 1981; Siegel, 1990; Young, 1982) written composition tasks (e.g., Flower & Hayes, 1981; Hayes & Flower, 1983; Hayes, Flower, Schriver, Stratman & Carey, 1987), personality psychology (Panter, 1990; Popham, 1996) and psychiatry (Yamauchi, Ono, Baba & Ikegami, 2001), an extensive literature search on cognitive process models for rating tasks to date has not resulted in any articles on the subject. The following section outlines the need to explore and model these processes during the task of rating.

Rationale

While TA protocols have been applied in educational and health research, there has been limited use of this qualitative method with the aim of obtaining rich, in-depth information about the cognitive processes underlying the task of rating in general. Moreover, to the best of our knowledge, the TA method has not been applied to the task of rating written policy documentation using a coding scheme; nor has a cognitive model of the involved processes been developed. In the area of tobacco research, questions often arise such as: What does this policy score mean? How did it come about? Should we feel confident using these scores as predictors of smoking outcome? It was felt that by far the richest source of information about rating would be to observe how the expert raters actually generated a response. This is important if one intends to assess the validity of inferences made from rating data obtained through a policy rating task. The interest in examining raters' cognitive processes during the policy rating task arose from the following practical issues and needs in the area of tobacco research:

1. Researchers construct quantitative indicators from tobacco policy ratings using the Stephens & English (2002) policy rating rubric.
2. Statisticians use these indicators in models to predict smoking outcomes and to make policy decisions affecting Canadian schools in all provinces.
3. There is researcher interest (University of British Columbia, University of Waterloo) to construct a new self-assessment tool for Canadian schools to assess their tobacco policies, for which the Stephens & English rubric is the basis.
4. Applied tobacco researchers and policy decision makers require information about validity evidence of inferences from ratings based on the Stephens & English policy rating rubric.

It is due to these practically salient potential consequences for policy that the validity of tobacco policy rating scores needs to be systematically addressed. Further, these issues are critical for all assessment, since validity, reliability, comparability and fairness are more than just principles of measurement; they are social values that have impact whenever an evaluative judgment is made (Messick, 1994). This emphasizes once more that validity, assuming both a scientific and a social role cannot be fulfilled by a simple correlation coefficient between test score and criterion. Hence, the need was to generate constructive input to potentially revise the rating process for future work. Currently and to the best of our knowledge, no such analysis into the validity of inferences made from this type of rating process is available in the literature, let alone using a TA method. The present research aims to fill this gap. To understand rating, one needs to identify the processes involved – but this would not be sufficient. One also needs to understand how raters organize these processes to produce the rating. Specifically, one needs to know how response processes are sequenced and related, how some processes are interrupted by other processes or terminated and how raters detect and correct errors. It is also important to know how simultaneous processes interact. In short, one would be interested in a model showing the processes involved in tobacco policy rating and accurately describe their organization and interactions. Such a model would serve as a metaphor for the process (Hayes & Flower, 1983). It would also serve the purpose of helping researchers speak to some critical validity questions in the field of tobacco research and help them see things in a way they did not see them before with respect to policy score meaning.

Today's researchers are interested in understanding the processes underlying participants' responses to tasks and problems in a wide range of areas. TA studies are a useful tool to obtain rich verbal data about reasoning processes. Moreover, using TA and subsequent protocol analysis enables researchers to identify the information that participants focused on while problem solving and how they used this information to facilitate solutions. The following section will describe the TA method, its strengths and potential limitations.

The Think-Aloud Method

TA studies have been widely used to study cognitive processes in many areas of psychology and education. For example, TA protocols have been used extensively in educational and descriptive research to assess academic and practical problem solving skills (e.g., Aanstoos, 1983; Banning, 2008; Bartolone, 2004; Berne, 2004; Block & Israel, 2004; Cumming, 1990; Davey, 1983; DeRemer, 1998; Ghaith & Obeid, 2004; Wilhelm, 2001; Carpenter et al., 1989; Lucas & Ball, 2005; Montgomery & Svenson, 1989; Phelps, 1990; Shapiro, 1994; Sienot, 1997; Wedman, Wedman, & Folger, 1996; Van Den Haak, De Jong & Schellens, 2007; Yang, 2003) and reading comprehension processes (Ericsson & Simon, 1993; Olson, Duffy & Mack, 1984; Pressley & Afferbach, 1995; Meyers, Lytle, Palladino & Devenpeck, 1990). In addition, TA studies have been applied in health and nursing research to understand cognitive processes in expert decisions making tasks (e.g., Backlund, Skaner, Montgomery, Bring & Strender, 2003; Funkesson, Anbäcken, & Ek, 2007; Jaspers, Steen, Bos & Geenen, 2004), as well as in health psychology to shed light on respondents' thinking while answering theory of planned behaviour questionnaires (French, Cooke, Mclean, Williams & Sutton, 2007), and in human resource management (Heerkens & Van Der Heijden, 2005). However, TA applications can easily extent to other contexts such as raters applying codes to written policy documents. To enhance understanding of the processes and issues involved in obtaining the TA data, the following section will provide some background on the TA method, along with a brief look at its advantages and limitations.

The TA method, in line with other methods that produce qualitative data, seeks rich, in depth-data from a small sample. The value of TA data rests on assumptions about verbalized data (Ericson & Simon, 1987). These assumptions are that a) cognitive processes that generate verbalizations are a subset of cognitive processes that generate any type of recordable response or behaviour, b) human cognition is information processing, and as such, a sequence of internal states successively transformed by a series of information processes, and c) information recently acquired and currently being concentrated on is directly accessible as verbal data using the TA method (Ericson & Simon, 1987). It is the last of these assumptions that may be viewed with skepticism, since it would be impossible to truly know the extent to which the cognitive contents are indeed expressed by the respondent. In addition, how could one ascertain that the verbal data so generated are an accurate reflection of the cognitive contents in the respondent? To put the skepticism towards TA studies into perspective, the next section will address some limitations of this method.

Potential Limitations of the TA Method

Historically, researchers have viewed TA studies with skepticism; this is due to the possibility of generating inaccurate data from verbal reports. This skepticism has been based on the assertion that the machinery of thought processes is inaccessible and hence cannot be captured as data (Nisbett & Wilson, 1977; Miller, 1962). A commonly acknowledged limitation of TA studies has been inconsistencies in data collection and the inability to verify findings from protocol analysis. However, this limitation can be addressed by following specific procedures to obtain accurate verbal data and analyzing data in a standardized step-by step manner (Fonteyn, Kuipers & Grobe, 1993). In addition, one needs to distinguish two types of verbal reports when obtaining TA data: concurrent and retrospective (Ericson & Simon, 1987; Van Den Haak, 2003).

Retrospective reports aim to reflect thought processes in the past; concurrent verbal reports arise from researchers' instructions to participants to 'think aloud' or 'talk aloud' while performing a problem solving task. This report type provides direct verbalizations of current cognitive processes and hence, is believed by some to be consistent and complete (Fonteyn et al., 1993).

It is acknowledged that due to the nature of human cognition, data obtained through a TA may remain somewhat incomplete. That is, thought is non-oral in its form and hence, can proceed much more rapidly than speech. When a series of thoughts occurs rapidly, it is impossible for an individual to verbalize each and every thought in that series. Since there is no way to analyze unreported information, no conclusions can be drawn or inferences made with respect to thought that occurred but was not reported verbally. To counteract these potential limitations, every effort needs to be made to remind subjects to "keep thinking aloud" after a few seconds (approximately 5-10 seconds) pause, then the percentage of unreported information should be kept to a minimum, compared to what *is* reported. Hence, the researcher obtains the fullest possible description of the reasoning used during the task (Fonteyn et al., 1993).

Other skepticism towards the TA method revolves around the notion that almost all of the subject's conscious effort is directed at solving the problem and hence, no cognitive resources are left for reflecting on what he or she is doing. However, as Ericsson & Simon (1993) discuss and show, talking aloud does not generally interfere with task performance. That is, thinking aloud is an activity which, in principle, does not cause much disturbance of the cognitive process. The subject performs a task while thinking aloud; it is executed almost automatically. Hence, TA data are very direct, there is no delay. Subjects do not give an interpretation of their thoughts nor are they required to put them into a predefined form; they just render them as they come to mind. Therefore, compared with structured elicitation methods, TA makes it easy for participants since they are allowed to use their own language (van Someren, Barnard & Sandberg, 1994).

Distinction of the TA Method from Introspection

As pointed out earlier, the TA method is still regarded with some methodological skepticism. This skepticism towards the TA method stems from the historical fact that the TA method was developed as an alternative to introspectionism, a subjective method by which subjects would act as detached observers of their own mental activity and expert witnesses of their consciousness. After the rise of behaviourist psychology, all first -person descriptions became a tabu. However, this radical eradication also meant the loss of information through alternative descriptive approaches.

It had been through the Gestalt psychologists that the descriptive method was utilized in problem solving and that the TA method was used. The Gestalt psychologists Wertheimer (1945) and Duncker (1926) exemplified the Gestalt contribution of being able to focus more holistically on the experience of thinking than had previous descriptive methods. Methodologically, their contribution was to make a clear distinction of how to use the TA method differently from introspection. According to Duncker, the introspector makes him or herself a thinking object of his/her attention; the subject who is thinking aloud remains immediately directed to the problem, so to speak allowing his/her activity to become verbal (Duncker, 1945). This important distinction was overlooked when all first-person description methods were discarded from psychological research on account of introspection’s failure. As emphasized in more recent research (van Someren et al., 1994), the suspicion towards the TA method as being too much like introspection is not justified for two main reasons: i) TA avoids interpretation by the subject and only assumes a very simple verbalization process and ii) TA treats the verbal protocols that are accessible to anyone, as data thus creating an objective method. To summarize, while it is healthy to consider potential limitations of the TA method, it is also important to gain perspective on the nature of criticisms expressed toward the method; this way, one can acknowledge the methodological significance of qualitative data obtained through TA protocols and ensure the necessary methodological rigor to minimize errors. The following section will provide more information on the methodological significance of TA data.

Methodological Significance of TA Data

The methodological significance of data gathered by having participants thinking aloud needs to be clarified to remove misconceptions of the meaning of such data that arose from both instrumentalist behaviourism and representationalistic cognitivism. Once these misunderstandings are removed, TA verbalizations can be gathered and analyzed properly. These verbal data can provide the researcher with valuable information on the process of responding - information that is not obtainable from purely quantitative studies that aim to illuminate relationships among variables; these studies cannot answer questions as to how these data arose. Hence, data from TA studies serve the important function of complementing quantitative psychometric methodologies and have the added bonus of illuminating the nature of one’s data by describing the underlying cognitive processes of responding during tasks. Leading authors on validity (e.g., Zumbo, 2005; Zumbo, 2009; Kane, 2001; Kane, 2006) have emphasized the need for an understanding of these processes that generates respondents’ answers. Hence, the TA method is an indispensible tool for those interested in an explanation-focused approach to validity (Zumbo, 2009) in their research. TA studies are not only valuable for describing what kind of information is concentrated on, but also how information is structured and organized during the task. This allows making inferences about thought processes applied during the task, and via this route, gathering evidence for the validity of inferences one aims to make from those responses. To clarify this advantage further, the next section briefly reviews the use of TA methodologies and protocol analysis and describes these in the specific context of rating tasks.

Protocol Analysis and the Specific Use of TA for Rating Tasks

In the area of language assessment, Lumley (2002) used the TA method to study how raters of written composition interpret and use rating scales. In Lumley's study, four trained raters of a high-stakes test marked 24 scripts and were asked to think aloud for 12 of these scripts. The study showed that TA protocols are useful in that they impart better understanding of cognitions in raters when they engaged in applying a rating scale to written composition. One can extent this concept to the task of rating written policy documents.

Studies of cognitive processes in medical decisions have mainly concerned tasks of diagnosis (e.g., Elstein, Schulman & Sprafka, 1978; Kuipers & Kassirer, 1984; Kuipers, Moskowitz & Kassirer, 1988; Moskowitz, Kuiper & Kassirer, 1988) and choosing between two or more alternative actions (Backlund et al. 2003). Backlund et al. (2003) compared the validity of rating scales and of TA protocols in a medical decision task (treatment of high cholesterol), exposing 20 doctors to six case vignettes. Doctors were asked to think aloud and ten of them were also asked to rate their inclination to drug prescription during successive phases of the decision making process. The TA data were found to be more sensitive to the directionality of the decision process. The authors concluded that the results generally supported the validity of TA data and rating scales as descriptors of decision processes.

The objectives of the present study were to identify and describe categories and themes about cognitive processes of rater responding while coding tobacco policies, describe how these processes are structured and organized, identify and describe obstacles raters encounter during the rating task, and identify and describe what raters do to overcome these obstacles. To meet these objectives, the study addressed the following research questions:

1. What cognitive processes are involved in the rating process, with raters using the Stephens & English (2002) rubric?
2. How are the processes organized?
3. How do these cognitive processes of policy rating contribute to the substantive aspect of construct validity via enhanced understanding of the measure and rating score meaning?
4. What obstacles distract raters during the rating process?
5. What cognitive processes do raters engage in to overcome these obstacles and to complete the rating task?

Methods

For the following section, methodological guidelines and criteria for reporting qualitative research were used. This was done for the purpose of working in a systematic manner, thus enhancing transparency of the TA study every step along the way. Specifically, the consolidated criteria for reporting qualitative research (COREQ) were adapted from Tong, Sainsbury & Craig (2007). The methodology described in this section is rarely done in construct validation and to our knowledge there are no models for researchers to follow. Therefore I describe this methodology in detail so that it may serve future researchers as a starting point or as an exemplar.

Research team, reflexivity to participants and study context

The first author conducted the TA protocol. At the time of the study, the first author was employed as a statistician at the Tobacco Research Unit (TRU) at the University of British Columbia and immersed in tobacco prevention research in the same team the raters worked in. The researcher's relationship with the study participants (raters) was collegial, in that all involved were working on common projects at the TRU during the time of the study. Participants were generally cognizant of the researcher's overall goals and reasons for conducting the TA study.

The present TA study of raters is situated within the context of a larger project, the Youth Smoking Survey (YSS) during the 2006/07 school year. The YSS is a national study of youth smoking in Canadian schools. The YSS study involved the same two expert raters as in the present study; they rated a total of 196 school tobacco policies each approximately four months prior to the present study. Hence, the raters had seen these policies before. However, since there was a substantial time gap between rating policies within the YSS and the present TA study, memory and recency effects are highly unlikely.

Theoretical framework: Content analysis

Content analysis was used in order to derive meaningful categories and themes describing the cognitive processes of raters and clarify the interrelationship among these themes. The aim was to understand participants' cognitive categories and to see how raters use these in the concrete task of tobacco policy rating using items of the Stephens & English (2002) coding rubric. Within the content analysis, the research method employed was protocol analysis, with two trained and experienced raters being asked to think aloud while rating. The content analysis was informed by grounded theory, in that the concepts of open coding and category development (Strauss & Corbin, 1990) as well as the constant comparative method (Glaser, 1965) were applied. In addition, concepts from discourse analysis (Rapley, 2007) were used. However, these methods only informed the content analysis, which was the primary method of analysis. The next section will briefly describe content analysis as a technique.

Broadly, content analysis is "any technique for making inferences by systematically and objectively identifying special characteristics of messages" (Holsti, 1968, p. 608). Objective analysis of messages conveyed in the data is accomplished by means of explicit rules, or criteria of selection specified before the analysis (Berg, 2007). The criteria of selection must be sufficiently exhaustive to account for each variation of text content and must be consistently applied so that other researchers examining the same data using the same criteria would obtain comparable results. This would be considered part of the research quality aspect, in terms of reliability checks and validation of findings. The categories established in the process should reflect all relevant aspects of the messages and retain, to the degree possible, the exact wording used in the statements of participants' thoughts (Berg, 2007). One can create a series of tally sheets to determine frequencies of relevant categories. This allows researchers to examine cognitive processes, topics and themes, while grounding the examinations in the data.

Content analysis is unobtrusive and provides a means by which one can study a process that occurs over periods of time. Limitations include the issue of locating unobtrusive messages relevant to the question, and one cannot test any causal relationships between constructs. Hence, the researcher must resist temptation to infer such relationships, particularly when presenting proportions or frequencies with which he/she observed a theme or pattern in the data. It is emphasized that while this type of information is appropriate to indicate the magnitude of certain participant responses, it is not appropriate to attach cause to these presentations (Berg, 2007).

It is acknowledged here that content analysis is sometimes viewed as another form of quantitative analysis, since in content analysis, one can use counts of certain textual elements. Hence, the implication is that content analysis is a reductionist, more positivist approach. However, as Berg (2007) argues, content analysis can be very effectively applied as a qualitative method, in that counts of textual elements provide a means for identifying, organizing, indexing and retrieving data. Data analysis, in contrast, involves consideration of literal words in the transcript, including the manner in which these words are presented. In essence, this part involves data interpretation, developing ideas about the data found in the various categories. In turn, the analysis needs to be related to the literature, broader pragmatic concerns, context of the task and the original research questions. Hence, the analysis provides the researcher with a means of learning about participants' thought processes and how these relate and fit into a larger context or issue. From this perspective, then, content analysis is not reductionist; it is a tool for listening to the words of participants and understanding better the perspectives of the producer of these words (Berg. 2007). As such, content analysis can be a powerful tool for cognitive process modeling for the purpose of gathering validity evidence to support one's inferences from respondents' answers.

Strategic decisions

The present study proceeded from several strategic decisions about how to conduct this research. Briefly, these decisions were:

1. To focus on the act of tobacco policy rating
2. To work holistically, e.g., let the data speak for themselves rather than having pre-existing hypotheses
3. To conduct an overall analysis of tobacco policy rating (all types)
4. To conduct individual analyses according to the policy sampling frame developed in study one of this dissertation (Zeisser et al., 2009), to be able to speak to any differences in the rating process depending on the type of policy rated (e.g., school-developed or board-developed)

The first and most important decision was to focus on the process of policy rating – that is, to attend to whatever it is that raters do when they produce policy ratings. Thus, rating was viewed primarily as a process rather than a product.

Participants

Participants were two expert raters, one female and one male. Both raters were purposively selected for their common characteristics; they have significant tobacco policy rating expertise from working in the Tobacco Research Unit (TRU) at UBC. This was important because the interest was in understanding the cognitive processes in expert raters rather than people who are learning to rate policies. Each rater had received previous training on coding written tobacco policies and has extensive experience in this task, each having previously rated approximately 360 written tobacco policies within the context of the Youth Smoking Survey and Project Impact. Both raters work in the TRU in the School of Public and Population Health at the University of British Columbia and are intensely involved in a larger research project about tobacco prevention and the area of tobacco policy rating. As such, the nature of the relationship between participants and researcher was that of colleagues. The expert raters were approached face-to face for participation, and at this point already had detailed understanding of the task they were invited to participate in, since raters had been carrying out this task extensively in their regular work.

There was the practical resource limitation of having only these two expert raters available; however, the number of raters was also compared to recommendations from the literature emphasizing that methodologies to shed light on complex thought processes require "rich data about individuals rather than easily analyzed data about a population" (Kuipers & Kassirer, 1984, p. 365). Silverman (2006) and Sacks (1995) also recommend that if depth rather than breadth is the aim and to make qualitative analyses effective, it is imperative to have a limited and manageable body of data to work with. In addition, the number of raters was justified by comparing to the number of raters used in similar research that employed raters for a written task (Lumley, 2002), or other rating research in general (e.g., Weigle, 1994).

Data Sources and Sample

The number of policies for each rater was chosen to strike a balance between gathering sufficient verbalized information about a variety of policy documents and, on the other hand, obtaining high-quality data by preventing rater fatigue effects due to rating unnecessarily numerous policy documents. The policies were randomly selected according to strata identified in the smoking policy characterization study (Zeisser et al., 2009). That is, long and short policies, school and board policies as well as comprehensive or less comprehensive policies were sampled to ensure that the variety naturally occurring within Canadian tobacco policies was reflected in the TA protocol analysis. This was important since these characteristics could have an influence on cognitive processes raters engage in during the task. The problems to be addressed next are the policy sampling frame and the item sampling frame.

Policy sampling frame

Based on prior experience of the raters, it was estimated that it would take each rater roughly ten minutes to think aloud while rating one tobacco policy using the sampled rating items assigned to them. One policy was randomly sampled to represent each facet combination in the policy sampling frame, a total of 12 policies. Hence, it was expected that each rater would produce approximately 1.5 hours of TA protocol data to be transcribed and analyzed qualitatively. Table 2 displays the policy sampling frame for the TA study and the order in which raters rated the policies (in brackets).

Table 2. Think-Aloud Study Policy Sampling Frame and Order in Which Raters Coded Policies

Policy Types	Length	Comprehensiveness			
		1-2 Components (Minimal)	3 Components (Medium)	4-6 Components (Maximum)	Total
School	Short	1 (1)	1 (2)	1 (3)	3
	Long	1 (4)	1 (5)	1 (6)	3
Board	Short	1 (7)	1 (8)	1 (9)	3
	Long	1 (10)	1 (11)	1 (12)	3
	Total	4	4	4	12

Item sampling frame

In the context of the YSS 2006/07, both raters had already rated all 196 school tobacco policies. In this context, the inter-rater reliability was established for each item of the Stephens & English (2002) rubric. The inter-rater agreement over all items was 96.6%. Of the 95 items (rating tasks) originally comprising the full Stephens & English (2002) coding rubric, seven items with the lowest inter-rater agreement (e.g., 90% or less) were selected for use by both raters in the TA study, because those items would be likely to be more problematic for the raters. In addition, 13 items with good but not perfect inter-rater agreement (e.g., 90.1 – 94.9%) were randomly split so that rater A rated seven of these items and rater B rated the remaining six items. Hence, those Stephens & English items with the lowest inter-rater agreement were used by both raters, while the indices with good but less than perfect rater agreement were randomly split between the raters.

Rater fatigue is a serious issue in TA studies; it could compromise the quality and accuracy of the rating data. To avoid this methodological problem, each rater rated only one policy from each facet combination resulting from policy types, lengths and levels of policy comprehensiveness, and only a sample of Stephens & English (2002) rubric items. To see which items were rated by both raters and which were randomly assigned to either one or the other rater, please refer to Tables 3 and 4.

Table 3. Items with Inter-Rater Agreement of < 90 % Used by Both Raters

Item	Wording	Agreement (%)
Sch1.0	Where was the school policy adopted from?	84.0
Sch6.0	How should the tobacco policy be communicated to parents/guardians?	88.8
Sch8.1	Does the policy outline the intent of the policy?	88.2
Sch11.2	Does the policy prohibit tobacco use on all school grounds?	89.3
Sch12.0	Does the policy prohibit teacher smoking anywhere on school property?	84.0
Sch18.0	Does the policy specify that sanctions should get stronger with subsequent violations?	89.3
Sch22.0	Does the policy specify who should be doing the disciplining?	89.8

Table 4. Items with High (90-95%) Inter-Rater Agreement Randomly Assigned To Rater A or B

Item	Wording	Assigned to Rater for TA	Agreement (%)
Sch5.0	How should the tobacco policy be communicated to teachers/staff?	A	92.5
Sch8.2	Does the policy outline the rationale of the policy?	A	91.4
Sch9.0a	Is there a blanket statement that prohibits everyone from smoking?	B	93.6
Sch9.2	Does the policy prohibit use of tobacco by teachers?	A	94.7
Sch9.3	Does the policy prohibit use of tobacco by visitors/parents/guardians?	A	93.0
Sch10.3	Does the policy prohibit smokeless tobacco?	B	91.4
Sch11.4	Does the policy prohibit tobacco use in school buses or vehicles used to transport students?	A	94.1
Sch13.0	Does the policy prohibit tobacco use at all times?	B	91.4
Sch15.0	Does the policy prohibit possession of tobacco products?	A	93.0
Sch16.0	Does the policy prohibit possession of tobacco products?	A	94.7
Sch19.0	Does the policy prohibit students from wearing tobacco brand-name apparel?	B	92.0
Sch23.0	Does it specify in the policy that there is tobacco education available for students?	B	90.0
Sch24.0	Does it specify in the policy that access and/or referral to cessation programs is available to students?	B	93.6

Setting

To gather TA data, each rater was scheduled for an individual session in a quiet room in the TRU that facilitates thinking aloud. Raters had a large desk space available to sit comfortably and carry out the rating task. No one else was present during data collection besides the researcher and the participant. The researcher informed the raters that she would record the session and take notes; the researcher also ensured the raters that the TA and notes are solely for research purposes and would not in any way be used for evaluating their job performance. The researcher conducted the TA protocol with each rater individually to ensure privacy and allow each rater to freely verbalize about the task at hand.

Task Environment

The task environment is considered to contain everything outside a rater's skin that influences the performance on the task. Hence, the task environment included the policy documents, the sample items from the Stephens & English (2002) coding rubric, written instructions and sample policy rated for practice and policy characteristics. Once rating had begun, the task environment also included the ratings the rater has already produced. These materials are relevant because raters refer to them repeatedly during the process.

Procedure

The researcher read the instructions out loud to each rater and they also received this detailed instruction sheet about the task in writing. The task was essentially the same as their regular work of policy rating at the TRU. Each rater also received a fifteen-minute individual TA training session where (s)he had the opportunity to go through a practice run (separate from the policies under study) to raise any questions or concerns they may have about the task, and to ensure their comfort with the task. None of the experts raised any concerns about the task and both gave their informed consent.

Next, raters received the 12 tobacco policies to be rated and a scoring sheet containing the sample items of the rubric. The raters were asked to constantly think aloud while engaged in the task. Specifically, they were asked to describe the rating process they engaged in when coding tobacco polices. Raters were also instructed to reflect on which rating criteria were vague or ambiguous. If a rater paused for longer than ten seconds because he or she encountered difficulties with assigning a code on a criterion, or did not understand a passage in a policy, the researcher prompted the rater to continue thinking aloud (verbalizing) in order to capture precisely these thought processes. Aside from this reminder, all interaction between raters and investigator were kept to a minimum to avoid interference with raters' flow of thoughts. The expert raters were debriefed after the TA sessions.

All TAs were recorded using a digital voice recorder, as well as backed up on a secure data storage location in the researcher's office. In addition, field notes were taken during the TA about what was being verbalized, especially points that needed clarification. The duration of the first TA (rater A) was 60 minutes; the duration of the second TA (rater B) was 90 minutes. All recordings and notes are kept securely in a locked filing cabinet accessible only by the primary investigator.

Problem solving tasks

To aid in the formulation of the task instructions sheet, a list of tasks was initially established to cover the problem solving process of rating the policies. The entire set of tasks was as follows:

1. Read the items provided to you carefully.

2. Read each tobacco policy provided to you carefully.
3. Use the rating items to rate each tobacco policy according to each of the coding criteria (rating one policy document at a time).
4. Enter the code you decided to assign for each criterion on the sheet provided to you.
5. As you are assigning your ratings, please verbalize aloud your thoughts about this process and the rating you are assigning for the policy on each criterion. Do not skip ahead in the items if you encounter difficulties, but keep on verbalizing your thoughts about your rating process.

The detailed task information and verbatim instructions for thinking aloud that raters received prior to the task are shown in Figure 3.

Figure 3. Information provided to raters prior to the think-aloud task.

I am going to ask you to rate a sample of 12 written tobacco policies (from YSS 2006/07), just as you would during your work at TRU. First, you will rate one policy for practice to ensure you understand the nature of the task and feel comfortable. You may ask questions for clarification if needed, or raise any concerns you may have.

I would like you to rate the policies using the subset of items provided to you. Please record the policy ID on top of the coding sheet, as well as all your ratings in the same manner as you would in the spreadsheet for your regular work. I would like you to talk and think aloud as you rate each of these policies, while this voice recorder stores what you say. As you rate each policy, please vocalize your thoughts related to this task and explain why you assign the codes you give. Please be as specific as you can; there are no 'good' or 'bad' thoughts. It is important that you keep talking during your rating task, registering your thought processes all the time. If you are busy reading the policy, please indicate it to me so that I can understand what you are doing at that time.

In order to prevent lengthy pauses in your rating and thinking aloud, I will sit here in the room with you and will prompt you to keep thinking aloud if necessary. I will prompt you if you fall silent for more than 10 seconds (except for reading policies).

To reduce cognitive overload and ensure data quality, we are going to take at least one break per hour of think-aloud, or as needed.

Please be assured that all verbal data will be used solely for research purposes by the researcher only. All data will be stored in a secure place and will in no way be used to evaluate your performance.

The original recordings were kept as a record and transcribed word for word by an experienced transcriptionist. Transcript text was checked for errors by listening to the voice recordings once again, checking against typed transcripts. Several small typographical corrections were made that would not have affected the meaning of text per se.

Coding of Transcript Text and Category Development

The first author coded the transcripts. The TA data were coded using a qualitative technique of identifying themes and categories, also referred to as qualitative data analysis of the process (Lewins, Taylor & Gibbs, 2005). That is, the data were coded according to themes and categories, with each category expressing a criterion for distinguishing some verbalized data from other verbalized data. One can then compare data within each category and draw finer distinctions within each category, allowing for a more detailed comparison of data organized within subcategories (Dey, 1993). For example, potential categories could be derived from what the verbalized data relate to – the policies, the ratings, a rater behavior or something entirely outside of the policies themselves. Then, subcategories can be developed based on barriers such as difficulty of interpreting the coding criteria (rubric items), extent of ambiguity in coding criteria, and coding dilemmas arising due to a lack of options for the coders (e.g., a policy component such as intent of the policy does not fit any of the coding options provided in the rubric). Further, one can analyze data assigned to different category levels by comparing and interrelating to produce a deeper analysis (Dey, 1993).

As Dey points out, categories must be conceptually and empirically grounded; they must relate to an appropriate analytic context and be rooted in relevant empirical data. For the present study, in order to present perceptions and cognitive processes of responding of policy raters in the most forthright way, a greater empirical reliance for category development seemed justified.

That is, in order to ensure that the categories reflect the thought processes of raters rather than any preconceived notions about the data by the researcher, all categories were derived from the transcript data, rather than imposed in advance. Category development occurred through immersion in the transcripts, establishing a dialectic between categories and data through repeated interaction between them. The emphasis here was on a holistic approach, attempting to grasp basic themes or issues in the data by absorbing them as a whole rather than trying to analyze them word-by-word or line-by-line. This approach also allowed for the moving of the analysis in either direction; in the process, one can go towards more refined distinctions through sub-categorization or a more integrated approach by linking and integrating the middle-order or main categories (Dey, 1993).

First, each full TA transcript was stored separately on its own 'data card' or stack (a separate word document). The four super-categories described in the coding section served as a first coarse unit of analysis. In the first coding step, all transcripts were read carefully and then re-read. The goal was to look for patterns, categories and ideas; this is part of the conceptual data analysis. This initial scan was rather coarse, but the purpose was to provide the researcher with a feel for the text data (Thompson, 2002; Daly, 2007; Maykut & Morehouse, 1994). As already emphasized, categories were derived from the verbal data, rather than in advance, in order to ensure the full capture of TA content. However, the following questions guided the coding process with respect to identifying themes and categories (first coarse categorization):

1. What cognitive process are raters involved in during the task of rating written tobacco policies, using each item of the rubric?
2. What type of information are raters processing in their mind during the rating task?
3. How are raters making their decisions before assigning their ratings?

4. Are raters making absolute judgments about each policy they code, or are they making relative judgments comparing to policies previously rated?
5. What prevents raters from understanding the policies clearly?
6. Which criteria in the rating process frequently present difficulties to raters when they assign a value to a policy component on an item?
7. What cognitive processes are raters employing in order to overcome the barriers and arrive at a rating?

With these questions in mind, and asking whether or not potential categories provide a useful basis for distinguishing differences or similarities in the data, open coding was used. Open coding refers to the naming and categorizing of phenomena through close and unrestricted examination of the data. The central purpose of open coding is to open inquiry widely (Strauss & Corbin, 1990). Transcripts were studied to discover concepts and categories that fit the data. These concepts and categories are, at this point, entirely tentative. While using open coding, some basic guidelines were observed: 1) ask the data a specific and consistent set of questions, 2) frequently interrupt coding to write a theoretical note. During open coding the data were broken down into discrete parts in order to arrive at themes. A theme can be a simple sentence, a string of words with a subject and a predicate. Specifically, open coding involves “breaking down, examining, comparing, conceptualizing, and categorizing data" (Strauss & Corbin, 1990, p. 61), often in terms of properties and dimensions. The next section describes in more detail the unit of analysis as used in this study.

Segmenting the Protocol into Units of Analysis: ‘Phrasing’

Rather than coding transcripts line by line, the focus was on segments - meaningful units of thought that are clearly discernible. Research on language production shows that in talk, pauses usually mark the boundaries of segments (Ericsson & Simon, 1993). This concept was adapted here so that a ‘phrase’ means a sequence of rater utterances fitting naturally within one meaningful unit of thought (a coherent theme). This was particularly relevant for defining the smallest unit of analysis for the finer-grained categories. The transcript text and the voice recordings were used in conjunction as the most reliable method for discerning phrases. This ‘phrasing’ is also consistent with Joseph and Patel’s (1990) method, whereby micro-units are conceptualized as segments of verbalized thoughts with a specific meaning; a micro-unit ends whenever there is a change in meaning. To facilitate the micro-analysis, each main category was divided into syntactic units or segments, using the method described by Joseph and Patel (1990). Since single words taken out of context can provide a severely misleading picture of what the talk is about, the micro-unit of analysis was chosen to be a meaningful segment of talk containing one thought. The following section describes stages of coding and how a category hierarchy was identified from the transcript text.

Development of Super-Categories

After initial reading of the full transcripts, a number of categories had begun to emerge and the focus was on defining a small number of super-categories to facilitate coding and to house subsequently defined main- and subcategories. For this step, the coding unit was the entire transcript. Two pre-existing issues provided the framework for generating super-categories. The first issue was that of raters interacting with a written document; the second issue was the nature of the TA task itself. In the coding process, the focus was on extracting relevant passages with references to common themes, categories and interpretation issues. That is, these super-categories provide a general framework for distinguishing rater verbalizations by general types such as rater behaviours, or verbalizations about the materials in the task environment (e.g., policies and rating items). Since they represent the coarsest level of categorization, the super-categories lend themselves to interpretation of cognitive processes in conjunction with finer-grained categories of raters' processes that had already begun to crystallize upon first-and subsequent transcript readings. Hence, the next challenge was to develop a level of categorization that would link the general framework with the finer-grained categories already noted during readings. The goal was to provide maximum explanation by being able to work laterally within the hierarchy of category relationships. That is, one can stay at the coarser level of the super-categories for the most parsimonious explanation, or one can move across the structure for more detailed explanations of rater responding.

Development of Main-Categories

Next, the focus was on raters' cognitive processes during interpretation and use of each sample item of the policy rating tool, the Stephens & English (2002) tobacco policy coding rubric. This was done with the intention to assess raters' understanding and interpretation of these rating items. Hence, this category level was developed by coding both transcripts at the item level and paraphrasing (in up to ten words) the main cognitive process raters engaged in to rate the item. Specifically, the goal was to summarize the *one* major cognitive process occurring for the rater, enabling her/him to rate the item at hand with a 'yes' (policy contains desired information) or a 'no' (policy does not contain the information). The analysis question of interest for this step was: What exactly is the rater thinking through while using this item to produce a rating? It was theoretically possible to arrive at more than one main category per item to describe the cognitive process pertaining to this item. However, for the sake of clarity and simplicity, every attempt was made to specify one and only one major cognitive process. Since the aim was to understand that process as a whole, the chosen coding unit was each item of the Stephens &

English (2002) rubric. These main categories are housed under the super-categories as defined above.

Main categories can also be referred to as in vivo categories (Strauss & Corbin, 1990) since they are the literal terms the participants used themselves. These in vivo categories then represent the cognitive processes, which will explain to the researcher how the basic problem of the participants is processed or resolved (Berg, 2007). All data for each main category that was identified were then brought together and were examined again as a whole. This was done to isolate the meaning and to verify that indeed each one of these categories is best described by this one thought process (a theme). If one item did not fit that particular process, it was allocated to a different category and if necessary, a new category was opened until no further new information could be obtained (e.g., when the category was saturated with the available data).

Refinement into Subcategories

While the raters' process of thinking through the item is critical, it was also of interest to break down the major cognitive process captured in the main categories via a micro-analysis to further enhance understanding of this process, which in turn would impart better explanation of the rating scores. That is, such a micro-analysis may help clarify the process in more detail, or it could challenge the previously assigned main category. Contradictory cases could then be derived from the micro-analysis and focused on in more detail. Hence, main categories of the cognitive process were examined in more detail by assigning finer-grained sub-codes within each item which would better explain the main process.

Subcategories were developed to present a more fine-grained level for representation of cognitive processes. These categories were based on the smallest meaningful unit of thought, rather than on line-by line coding (as described earlier in the section about 'phrasing'). They were developed to provide deeper-level information above and beyond the more general information from the item-level categories.

To check the consistency in meaning of these sub-codes, as with the main categories, all statements of one such subcategory were again transferred onto one ‘data stack’ (e.g., a separate word document with a unique descriptive title), read and then re-read carefully. The intent was to ensure that all statements and segments of verbalized thought that were given the sub-code were indeed best described by this subcategory. The same process as with the main categories was applied for non-fitting statements; upon critical re-reading, any non-fitting data pieces were re-examined and either transferred to a more appropriate category or a new subcategory was created. After this consistency check, each data stack containing a subcategory was read again in order to arrive at the final, more fine-grained description. Materials sorted in this manner were examined to ensure they speak to the subcategory under investigation. The subcategories are subsumed under each of the major cognitive categories derived in the previous step for each rubric item. That is, each subcategory can serve as a ‘finer’ description of a main category. In refining categories, the constant comparative method (Glaser, 1965) was used, whereby for each new piece of data, one constantly compares within and between existing categories. The next section addresses how codes were utilized in the analysis process, and illustrates how relationships among categories were analyzed.

Analysis

While frequency counts appear to belong solely to the quantitative domain, they do have value in content analysis since they provide information in and of themselves. Frequency counts also ground the results in the data. That is, a cognitive process engaged by the rater over and over again obviously provides insights into the relevance of that process for solving the rating task. It can also provide valuable information about the item and about the interaction between raters, items and rating objects (policies), as well as about the characteristics of policies (e.g., school or board, short/long, level of policy detail). As such, overall frequency counts and counts by sampling frame were also generated because they were regarded useful as a preparatory step for analyzing relationships among categories.

On a cautionary note, clearly, frequency counts of categories alone and without further analyses would merely present a surface look at the data by giving an overview of the occurrence of data pieces cast into each category. Nevertheless, having a large number of data pieces in one particular category does suggest to the researcher where to look for patterns. Further, if many of these data pieces express similar issues that establish a pattern, the researcher is able to offer an idea of how strong this pattern is by describing its magnitude. Hence, to be considered a relevant theme, categories had to contain at least three occurrences of data pieces in total (between both raters) within them, following recent recommendations by Berg (2007). Hence, as a tool to help facilitate understanding of the associations between raters' cognitive process categories (e.g., flow of categories, what processes interrupt running rating processes, how categories are organized), frequencies were examined first.

Please note that the general recommendation by Berg (2007) to have at least three data pieces before establishing a category was relaxed for the purpose of separately analyzing rating processes for school-developed versus board-developed polices. For this purpose, even data pieces with just one frequency count were considered in order to use as much information as possible for delineating differences with respect to the different objects rated. Hence, cognitive process models for school-developed or board developed policy ratings can contain categories with observed frequency counts of less than three.

Barriers and Obstacles to Rating and Raters' Strategies

In addition, the intention was to identify and code cognitive themes that indicate obstacles encountered during the process, as well as themes that indicate how raters cope and overcome these obstacles. The transcripts and field notes were reviewed, writing down direct quotes deemed especially relevant to the process of rating and interpretation of rating criteria.

The next exploration was how raters' process categories are associated with barriers and how barriers are associated with raters' coping strategies. This was accomplished by closely following the transcript sequences for each Stephens & English item rated. In analyzing raters' cognitive processes, obstacles encountered and strategies to overcome the obstacles, the focus was on the following actions and indicators of disruption shown in the transcript (Rapley, 2007):

- Delays: a gap before a response or a gap within a response, a delay before an answer is given.
- Hesitations: utterances such as 'mm', 'erm', 'uhm' and in-breaths or out-breaths.
- Prefaces: Phrases such as 'well', and 'uh' or agreement tokens such as 'Yeah'.
- Mitigations: apologies or appreciations.
- Accounts: Excuses, explanations, justifications and reasons.

To illuminate the nature of these obstacles to rating, key pieces of transcript pertaining to these occurrences were extracted as excerpts and examined more closely. If at all possible, the same was done for any actions or indications in the transcript text pointing towards raters' strategies for overcoming the obstacle (e.g., referring back to previous ratings or relying heavily on memory from similar obstacles previously encountered).

The Issue of Rigor and Safeguards Against Potential Flaws in the Content Analysis

To accomplish a content analysis of high quality, it was necessary to address how scientific rigor was to be accomplished and how analysis flaws can be avoided. Several steps were taken to avoid flaws (Berg, 2007). First, every assertion about the overall ratings was documented with at least three examples from the data. Second, analytic interpretations were examined carefully by an independent reader to ensure that these claims and assertions do not stem from a misreading of the data and that they have been documented adequately. Finally, whenever inconsistencies emerged, these were discussed with a second reader and resolved by using clearer wording and category re-assignment. In summary, the following steps were followed to ensure rigor and quality of the analysis:

- Describe how materials were generated, worked with and analyzed.
- Check and re-check ideas and findings against the data and search for instances that might refute or contradict claims made (through constant comparative method).
- For all central analytic points, provide the reader with detailed access to the data that led to those claims (include transcript excerpts or quotes).
- Keep a running research diary of analysis notes.

The following section provides details on the reliability check conducted with a second coder on a subset of transcript text.

Reliability Check on Categories

As Appleton (1995) points out, credibility in qualitative research can be enhanced when the researcher leaves a clear decision trail concerning the study steps. One crucial requirement for rigor is that the categories are sufficiently exhaustive and yet precise enough to enable different coders to arrive at similar results using the same data. Therefore, a second reader coded approximately 10% of the transcript to determine whether the codes are clear, that there are no major errors in the codes, that no major codes were missed and that it is reasonable to use these codes. The second coder was not a trained tobacco policy rater and was unfamiliar with the research on tobacco policies. However, the second coder was a colleague of the first author and a fellow graduate student in the field of measurement and evaluation. Hence, the second coder had a good understanding of the methodological issues around establishing rigor in qualitative analyses. The second coder was asked the following questions about the list of categories:

1. Do these categories make sense to you?
2. Is the list sufficiently refined, or are additional categories needed?
3. Can you think of better, more precise wording for the categories?
4. Are there redundant categories?

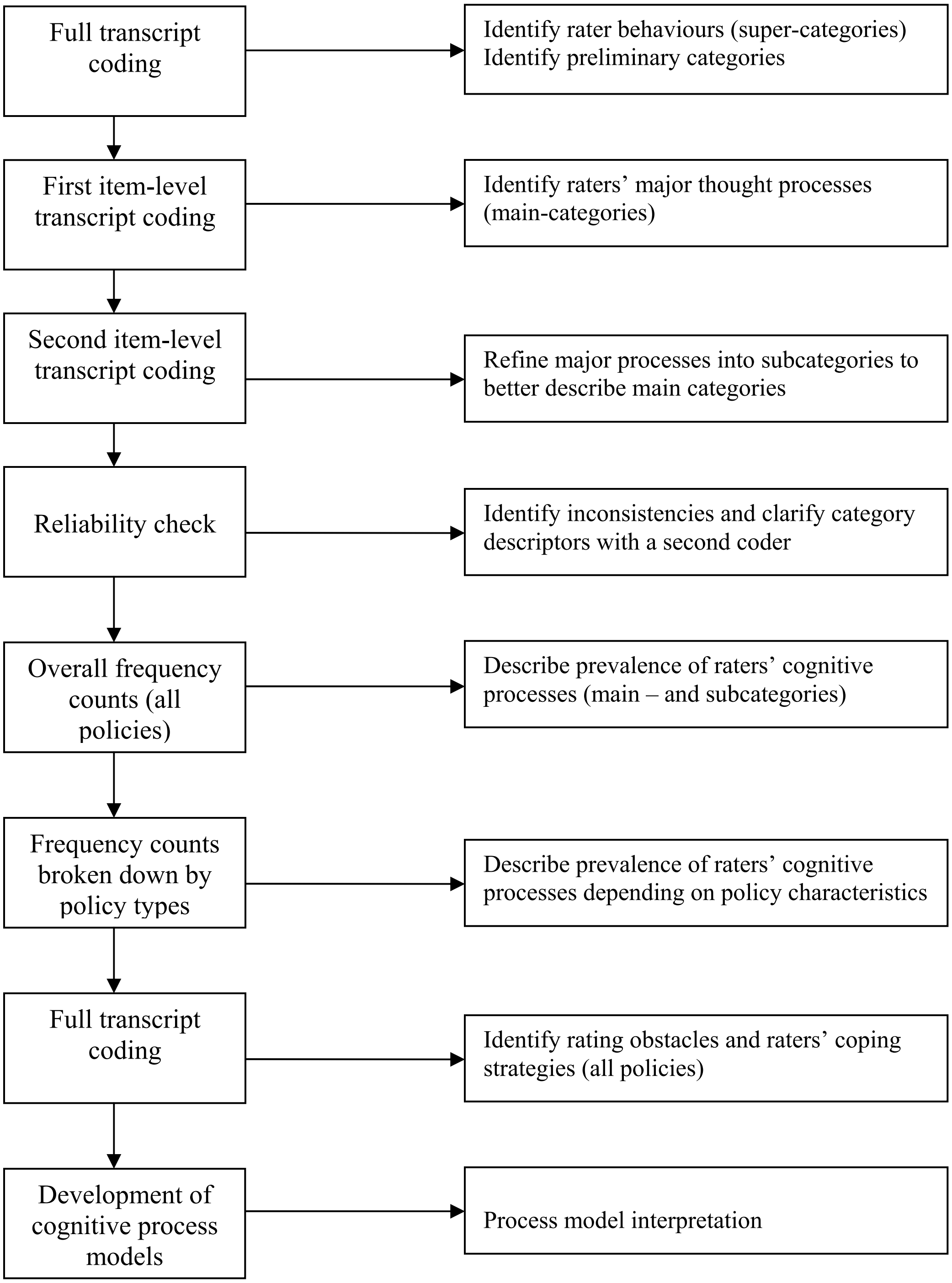
In addition, it was of interest to see to what degree the second coder arrived at the same main categories, using the same data. The reliability check revealed that there was 81% agreement between the two coders. For items where coders disagreed, in-depth discussions covering the set of question above lead to the following improvements:

1. Adjustments in wording for clarification in two category names.
2. Several instances of one troublesome main category that was too broad (5 – Arriving at a decision by integrating multiple pieces of policy information) were resolved by

subsuming under existing, better-defined categories (e.g., 14 – Deciding by reflecting back on one's general understanding of the content of the policy just read, and 12 – Searching policy for key information).

Figure 4 displays the flow of steps during the data analysis.

Figure 4. Flow diagram of data analysis.


Full transcript coding
Identify rater behaviours (super-categories)
Identify preliminary categories
First item-level transcript coding
Identify raters' major thought processes (main-categories)
Second item-level transcript coding
Refine major processes into subcategories to better describe main categories
Reliability check
Identify inconsistencies and clarify category descriptors with a second coder
Overall frequency counts (all policies)
Describe prevalence of raters' cognitive processes (main – and subcategories)
Frequency counts broken down by policy types
Describe prevalence of raters' cognitive processes depending on policy characteristics
Full transcript coding
Identify rating obstacles and raters' coping strategies (all policies)
Development of cognitive process models
Process model interpretation

Results

This section is structured in the following order: (i) description of cognitive process-categories and their hierarchy, (ii) presentation of overall category frequencies and (iii) frequencies broken down by the policy sampling frame (e.g. policy type, length and comprehensiveness).

Super-categories

The following super-categories were derived upon first and second read of the full transcript text: a) TA verbalizations pertaining to raters' behaviours (e.g., rater engages in information seeking), b) TAs pertaining to the ratings as such and the use of the rubric items (e.g., correcting one's own thought process or rating on a particular item), c) TAs pertaining to rating objects - the policy document currently rated (e.g., inferring meaning from a policy statement) and d) 'meta-TAs' – those TAs outside of the policy currently rated and outside of the item currently used to rate (e.g., expressing a personal opinion, using expert knowledge).

Categories for Main Cognitive Processes at the Item-level

A total of 12 major cognitive processes (item-level categories) were identified from the transcripts and they are described as follows:

- Confirming a rating decision by double-checking the policy.
- Deciding by asking and answering a question to oneself.
- Decision through a generalization (e.g., over people, places, concepts).
- Stating key information from policy (one piece sufficient) to justify a decision.
- Deciding by comparing to known standards of what makes a strong policy.
- Stating that key information to give a 'yes' rating is missing.
- Arriving at a decision by clarifying terms and meanings.

- Making a decision through a distinction (e.g., between concepts, people, places or products).
- Correcting his/her reasoning when coming across more information *after* a rating.
- Searching policy for specific information.
- Inferring that a criterion is met my listing several (≥2) pieces of information required.
- Reflecting back at his/her understanding of policy content just read.

On occasions, raters answered two adjacent and closely related items with one main cognitive process (e.g., rater A, C5: 'How should [the policy] be communicated to teachers? It only says that the principal is to inform them, so it doesn't specify how to inform them: so I'm gonna say no, and no for parents'). In such instances, the two subsequent items are so closely related in content that raters can answer them virtually at the same time engaging in the same cognitive process. As a result, it was possible to technically subsume two rating items under the same category, rather than counting this process twice. Therefore, the total number of cognitive main categories does not match the total number of items the raters rated.

After finalizing the main categories for cognitive processes occurring for each item, it was necessary to go back to the fine-grained codes that had begun to emerge during previous transcript readings. Transcripts were read again item-by-item, and the subcategories finalized as the following section describes.

Subcategories

Table 5 describes in detail the subcategories of rating and shows how they are housed under the super-categories (rater behaviours).

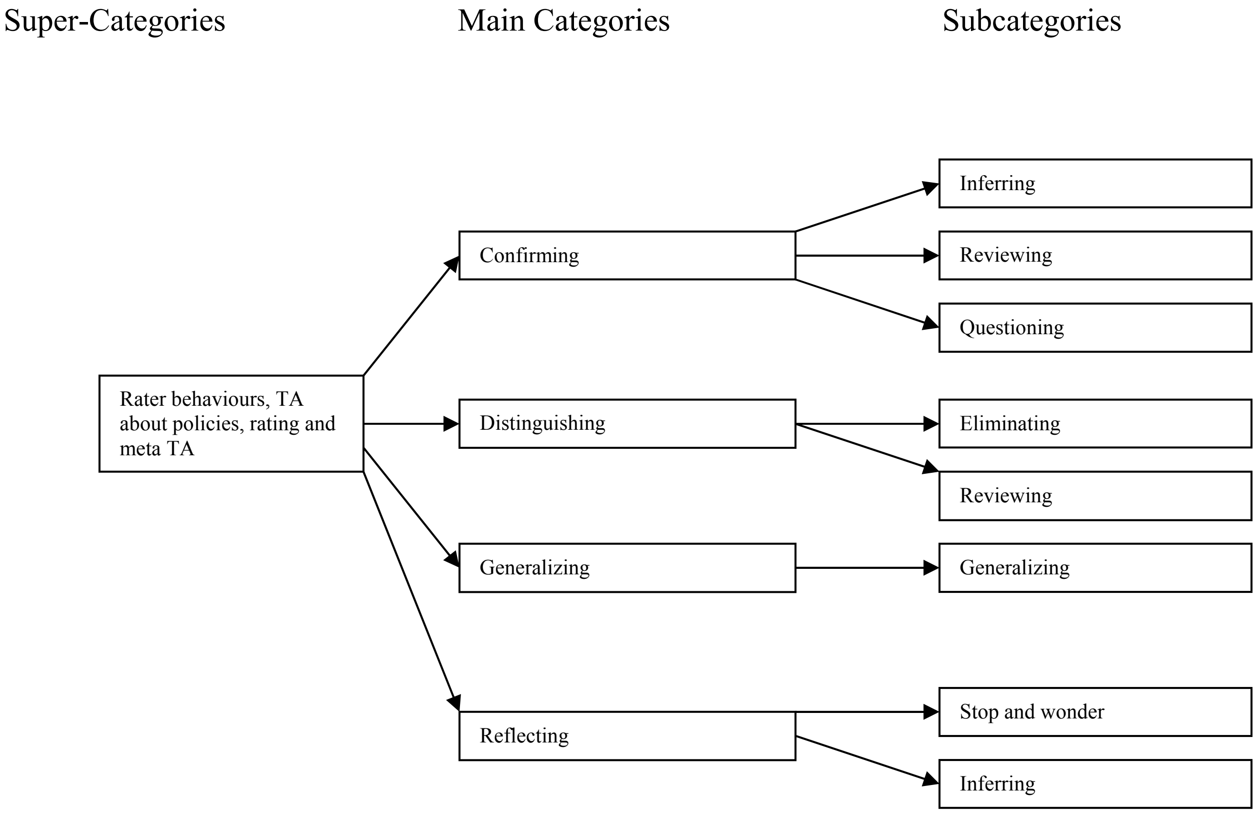
Table 5. Subcategories Description and Relation to Super-Categories

Subcategory	**Description and Relation to Super-Categories [a, b, c or d]**
Information seeking	Include any data where rater scans policy document (not re-reading) for information about an item and/or reads out the info, including stating that info not mentioned (=info not found because not there) [a]
Inferring	Include any data where rater decides by drawing meaning about a key policy element from other info in policy currently rated [c]
Rationalizing	Include any data where rater immediately provides logical reason(s) or justification for a rating decision (if x then y) [b]
Questioning	Include any data where rater asks questions to self for clarification of an item or policy content [b or c]
Checking	Include any data where rater double-checks his/her ratings if confused or unsure [b]
Eliminating	Include any data where rater decides by thinking about elements a policy does not specify but should specify to get a point [c]
Clarifying	Include any data where rater clarifies meanings of words and phrases in policy for him/herself [c]
Confirming	Include any data where rater integrates policy content to ensure it means what (s)he coded for [c]
Distinguishing	Include any data where rater differentiates the meaning of policy phrases that qualify for scoring a point from those that do not [c]
Excluding	Include any data where rater has to score 0 (‘no’) because the information is only available for a non-relevant issue but not the issue specifically asked in the item.
Contradiction finding	Include any data where rater points out contradictions in policy content [c]
Jogging memory	Include any data where rater tries to access memory from previous ratings and knowledge how to code [d]
Comparing to known standards	Include any data where raters state what a strong policy should contain and compare this to policy content pertaining to the item in question [d]
Reviewing	Double-checking or re-reading less than 3 lines of policy [a]
Using insider knowledge	Rater is able to make a decision by applying specific expert knowledge about his/her area [d]
Stop and wonder	Explicitly pausing and wondering [a]
Deciding on partial info	Only some of the required info is there but not all of it (policy says A but not B which would also be needed) [c]
Re-reading	Include any data where rater re-reads whole policy sections (> three lines) to make a decision if in doubt or states (s)he is re-reading sections [a]
Correcting	Include any data where rater corrects his/her thought process or rating [b]
Critiquing	Include any data where rater critiques policy for poor wording or lack of clarity [c]
Commenting	Include any data where rater expresses personal opinion or position [d]

Note: Please refer to descriptions of super-categories a, b, c and d at the beginning of the results section.

The literature reports the use of structure diagrams in qualitative studies involving category work (Dey, 1993) to show the interrelations among the themes and categories. Figure 5 displays such a diagram to show the category structure used in the present study. Please note that this diagram is not intended to show all categories (due to space limitations); it has the purpose of demonstrating the hierarchy of the categories and how they may relate to each other using examples, specifically how main categories fall into one of the super-categories, and how subcategories serve to better explain main categories.

Figure 5. Diagram demonstrating category structure using examples of categories.



Category Overlap

When developing finer-grained subcategories to explain main categories better, the observation was made that in some cases, only one subcategory could be found within the main category, and it was the same as the main category (e.g., main category at the item level = 'generalizing', subcategory = 'generalizing'). Hence, there was perfect overlap and the subcategory did not add any new information above and beyond the main category. In some instances, this was clearly due to very short TA responses to the item (e.g., rater A, B5: 'There's no intent given in this policy'). Hence, category saturation was reached at the level of the item. In other instances of complete overlap, there was no available information in the policy for the raters to TA about (e.g., main category at the item level = 'stating that key info is missing', subcategory = 'stating that key info is missing'). In instances where only one subcategory was found within a main category and they were identical, a subcategory code was not assigned and only the item level main category was counted. However, it was also possible to find within a main category a subcategory that serves as a main category for other items (e.g., 'distinguishing'). 'Distinguishing' is one of the main categories, but in some cases, a TA response at the item level is better described by another main category. Then, when fine-coding this TA response for subcategories, 'distinguishing' may be one among other subcategories that help better explain this response. Hence, it was possible to have main categories and subcategories with the same label.

Overall Frequency Counts

First, results for the use of the main cognitive process categories at the item level are shown in Table 6, displaying total frequencies (rater A and rater B). Examples are provided with excerpted key phrases verbalized by both raters to illustrate each main cognitive process. The overall frequency counts for use of the subcategories are displayed in Table 7, along with excerpted examples from both raters. Table 8 shows main cognitive processes in raters with total frequencies broken down by the policy sampling frame.

Table 6. Main Cognitive Processes in Raters: Categories at the Level of the Policy Rating Item with Total Frequencies

Main cognitive process category	**Excerpted key examples (rater A or B)**	**Freq. Total (%)**
Stating that key info to rate 'yes' is missing	…no, it does not mention sanctions or violations or consequences (B).	68 (30.9 %)
Stating key info to back a rating decision	… it says 'the division is a smoke free environment' that's a blanket statement, yes (B).	39 (17.7 %)
Making decision through distinction	Does the policy outline the intent of the policy?' No, all it does is list consequences (B).	38 (16.3 %)
Deciding via generalization	Yes. 'School property is a tobacco free and no smoking zone.' Uh that applies to everyone (B).	27 (12.3 %)
Deciding by integrating multiple pieces of policy info	...looking for zero tolerance, (3 second pause) what happens if I smoke on school property. Says 'if you are sixteen or older you be charged by Tobacco Enforcement Officer and be given a three hundred and five dollar ticket if you smoke on school property' (3 second pause) hmm so the Tobacco Enforcement Officer will… give you a fine the first time they see you…but that doesn't mean that if you get caught by the Principal that they're gonna report you to the Tobacco Enforcement Officer... Can I be, oh there's another section? 'Can I be charged for just holding a friend's cigarette and not actually smoking?' and it says 'yes you will be charged if you are holding a lit cigarette on school property.'... So I will say yes to zero tolerance. If you are caught you will, you will be charged and consequences there are consistent (B).	8 (3.6%)
Searching policy for specific key info	Now I'm gonna look back at the list of consequences...(B).	14 (6.4%)
Confirming decision by double-checking policy	That's an intent…well... 'To provide a working and learning environment that is free from tobacco smoke.' That's definitely intent. So yes (B).	8 (3.6%)
Inferring by listing two or more pieces required for decision	…it does mention consequences: It has first offence, second offence, so it does get stronger with subsequent violations (A).	8 (3.6 %)
Clarifying terms and meanings	Uh 'Widely accepted research has demonstrated the risks to health caused by second hand smoke.' That would be your rationale (B).	4 (1.8 %)

Main cognitive process category	**Excerpted key examples (rater A or B)**	**Freq. Total (%)**
Deciding by answering a question to self	Hmm… no that does not count huh (slight laugh). Uh and why not? Because it doesn't have anything to do with smoking (B).	3 (1.4 %)
Comparing to known standards	I'm gonna say no, um because a good policy should mention all times, it should say 'School property is a tobacco free and no smoking zone at all times' (B).	2 (.9 %)
Correcting one's reasoning by coming across more info	I'll choose yes - the policy prohibits tobacco use on all school grounds. Because the prohibition statement prohibits ….yeah. (3 second pause) oh no, that was the wrong tra', that was the wrong thought process hahaha, looking: OK now I'm looking at the uh at the question number, I know that this um, this question is not about the prohibition statement it's about the location (B).	2 (.9 %)
Reflecting back on one's understanding of policy content	Policy prohibits teachers from smoking anywhere on school property' it appears as though that's the case, so yes it is prohibited (A).	4 (1.8 %)
Total number of thought processes		*220 (100 %)*

Table 7. Subcategories of Raters' Cognitive Processes at the Unit of Thought-Level with Excerpt Examples and Total Frequencies

Subcategory	**Key Excerpt Example (rater A or B)**	**Freq. Total (%)**
Information seeking/scanning	Scanning policy (3 second pause). So the Human Resources Department will assume responsibility to advise all candidates for employment with the Division of the smoke-free environment policy.' ' It doesn't talk about advising or communicating other user groups. (B)	113 (28.1%)
Rationalizing	That does not count. Why not? Because it doesn't have anything to do with smoking. (B)	44 (10.9%)
Inferring	So that means outside of the school day they can smoke anywhere. (B)	40 (9.6%)
Clarifying	Rationale would be something like, because smoking is harmful to your health (B). Now they said that they want it to be a healthful environment…healthful environment, so I guess that that was an intent (A).	34 (8.5%)
Excluding	…this is a consequence for students so I don't know if it's prohibited for teachers, or parents, visitors, guardians (A). Does not prohibit teacher smoking, 'cos the policy does not mention teachers (B).	28 (7.0%)
Eliminating/rating based on absence of info	It does not mention the health of anyone, does not mention a goal of reducing smoking…	27 (6.7%)
Distinguishing within a policy	Doesn't have anything about possession so I'm 'no', it only mentions smoking (A). We have the question here about *intent* and there is also a question about *rationale*…(B).	23 (5.7%)
Confirming	It doesn't mention consequences…I don't think. No, it doesn't (A). Ahh, so first one is a blanket statement 'a non smoking ban shall exist in or on any property'. Yeah, that's a blanket (B).	15 (3.7%)
Re-reading if in doubt	And it doesn't mention consequences, so I'm just gonna read that section again (A). Now I'm gonna look back at the list of consequences (B).	14 (3.5%)
Questioning	What do I do in this case? (A)	12 (3.0)
Jogging one's memory	Um, rationale, I just need to refresh my mind the difference between intent and rationale (A). Uh, so in the other case it said…health and wellbeing of students…(B). I remember that 11.1. and 11.3 they are all about the places where smoking was prohibited (B).	8 (2.0%)
Expressing opinion	That one's a bit iffy (A). This was a really long policy (A). I think that's weak and I don't wanna give it to them (B). Well the language is pretty absolute (B). First time that I've seen such a specific policy today (B).	8 (2.0%)
Pause and wonder	…that sorta makes me stop and wonder…school property is tobacco free and no smoking zone	6 (1.5%)

Subcategory	**Key Excerpt Example (rater A or B)**	**Freq. Total (%)**
	(A). I'm wondering again: does this blanket statement apply to teachers (B)?	
Deciding on partial info	It does mention consequences but it only mentions one consequence: three day suspension, so that does not mention anything about um subsequent violations (A). It only mentions smoke-free environments inside the building at all times but that's not the extent of the policy (B).	6 (1.5%)
Comparing to standards	I would say no to it though because generally, it has to mention possession, be more specific o possession (A). …hmm, if they had mentioned the amount of the fines of the tickets then I could have said that they do get stronger (B). A good policy should mention all times, should say 'School property is a tobacco free and no smoking zone at all times' (B).	5 (1.3%)
Double-checking one's rating	Does not explicitly prohibit teachers smoking, but I did check yes for the blanket statement…(B). Cessation programmes – 'no', but I'm going to double check (B).	4 (1.0%)
Critiquing policy wording/clarity	Um it doesn't have much detail at all (A). Consideration is a bad word, that's not very strong (B).	4 (1.0%)
Using insider knowledge	But I think the allowing smoking in certain areas is more like for a smoke pit (A). The school is from Ontario and I know that they do (B).	4 (1.0%)
Reviewing/double-checking policy	Doesn't have anything about how to communicate…just double-checking, no (A).	3 (.7%)
Including	I may count that as a rationale, in this case, but there is no intent (B).	2 (.5%)
Contradiction-noticing	…up above it says to ban smoking and tobacco products on division property but in the actual policy, how the policy reads it's gonna say smoking and or the use of tobacco products…(A).	1 (.25%)
Correcting one's thought process	Because the prohibition statement prohibits…yeah. (3 second pause) oh no, that was the wrong thought process…(B).	1 (.25%)
Total number of thought processes		*402 (100%)*

Table 8. Main Cognitive Processes in Raters: Categories with Total Frequencies Broken Down by the Policy Sampling Frame

Main cognitive process category	**Freq. Total (%) School**	**Freq. Total (%) Board**	**Freq. Total (%) Short**	**Freq. Total (%) Long**	**Freq. Total (%) Min**	**Freq. Total (%) Med**	**Freq. Total (%) Max**
Stating that key info is missing	41 (35.7)	31 (24.7)	34 (35.4)	28 (26.2)	21 (31.8)	25 (33.3)	21 (29.2)
Stating key info to back a rating decision	17 (14.8)	20 (17.7)	14 14.6)	21 (19.6)	7 (10.6)	17 (22.7)	13 (18.1)
Making decision through distinction	17 (14.8)	17 (15.0)	17 (17.7)	17 (15.9)	6 (9.1)	12 (16.0)	14 (19.4)
Deciding via generalization	19 (16.5)	12 (10.6)	15 (15.6)	10 (9.3)	14 (21.2)	10 (13.3)	3 (4.2)
Deciding by integrating multiple pieces of policy info	9 (7.8)	12 (10.6)	2 (2.1)	14 (13.1)	7 (10.6)	5 (6.7)	5 (6.9)
Searching policy for specific key info	5 (4.3)	6 (5.3)	6 (6.3)	7 (6.5)	5 (7.6)	1 (1.3)	7 (9.7)
Confirming decision by double-checking policy	3 (2.6)	4 (3.5)	1 (1.0)	6 (5.6)	3 (4.5)	1 (1.3)	4 (5.6)
Inferring by listing two or more pieces required for decision	1 (.9)	2 (1.8)	2 (2.1)	1 (.9)	0	0	4 (5.6)
Clarifying terms and meanings	1 (.9)	2 (1.8)	1 (1.0)	2 (1.9)	0	2 (2.7)	1 (1.4)
Deciding by answering a question to self	0	0	2 (2.1)	1 (.9)	1 (1.5)	1 (1.3)	0
Comparing to known standards	1 (.9)	6 (5.3)	1 (1.0)	0	1 (1.5)	0	0
Correcting one’s reasoning by coming across more info	1 (.9)	0	1 (1.0)	0	0	1 (1.3)	0
Reflecting back on one’s understanding of policy content	0	1 (.9)	0	0	1 (1.5)	0	0
Total number of thought processes by grid	*115*	*113*	*96*	*107*	*66*	*75*	*72*

Main Obstacles to Rating and Raters' Coping Strategies

Key Excerpts Rater A.

This section presents findings about key obstacles that raters encountered, shows detailed examples and illustrates how raters overcame these obstacles to produce a rating. In some examples, raters could not find a coping strategy due to lack of information and only the rating obstacle is shown. Please note that rating obstacles are presented in regular font and raters' coping strategies are in *italic* font. The brackets after each excerpt example contain a capital letter indicating whether it was rater A or rater B, and an item code consisting of a capital letter and a number designating the location of the data excerpt in the transcript.

Excerpt 1….um [4 second pause] that one's a bit iffy 'cos it does say 'tobacco free' and 'no smoking zone' so no smoking means you can't ...actually smoke, but then the tobacco free…. possession… that would imply that you can't have tobacco, huh (exhales sharply as if to laugh). Um… 'Prohibits possession of tobacco products' [3 second pause] I would normally say that it didn't prohibit possession but because it distinguishes between tobacco free and no smoking that's sorta makes me stop and wonder…'*School property is a tobacco free and no smoking zone' t-t-t (clicks her tongue) (4 second pause) I'm still gonna, I'm gonna say no to that one* (A, B10).

Excerpt 2…*and* I don't know about all grounds, it just says smoking um. What do I do in this case? *'Policy prohibits tobacco use on all school grounds' (pages turn) I'm gonna say no it doesn't because it doesn't… it just says you can't smoke (pages turn). It says 'Smoking, policy prohibits tobacco use in all school grounds' (10 second pause) I would say no, it doesn't mention the grounds, I don't know where the smoking's prohibited.* (A, D7).

Excerpt 3. Um it doesn't have much detail at all, so 'intention' 'rationale' are not present in this policy. (A, J6).

Excerpt 4*…yeah*. The only reason I pause is it says 'leased facilities' *but facilities isn't vehicles but I think school buses are…school buses are separate so 'no'.* (A, K8).

Excerpt 5....uh this is, I see, I forget why we used this negative one, 'allows smoking in certain areas,' (3 second pause) *because it doesn't prohibit anything on school property, it's only in the school so: policy doesn't prohibit anywhere on property. I'm just trying to decide whether it's - no they don't prohibit it or that they allow smoking in certain areas. But I think the allowing smoking in certain areas is more like for a smoke pit... if it doesn't, if there's nowhere on school property that allows it …I think it would be: 'Policy prohibits teachers smoking anywhere on school property.' No it does not. It allows it which doesn't prohibit it. So I'm just gonna go with no on that one 'cos I don't think (3 second pause) I think the negative one means it's not really allowed but there's certain areas on school property where it is allowed.* (A, L10).

Key Excerpts Rater B.

Excerpt 1. I'm wondering - I'm wondering again: does the blanket statement apply to teachers? Uhh, for this item that asks specifically about teachers: 'Does the policy prohibit teachers from smoking anywhere on school property?' *does not explicitly prohibit teachers smoking, it says that school property is a tobacco free no smoking zone, which applies to everyone…uh so I'm gonna say 'no'.*(B, A10).

Excerpt 2. (8 second pause). Mm we have the question here about intent and there is also a question about rationale, and I'm thinking about what the differences were that we looked for*…intent is the goal and the rationale is the reason. Rationale would be something like: because smoking is harmful to your health and intent would be to promote a healthy lifestyle.*

Uh: "widely accepted research has demonstrated the risk to health caused by second hand smoke'. That would be your rationale. (B, B5).

Excerpt 3. (3 second pause). I get hung up about the words between the coding and the policy and I have to think about: well what's the important thing? Is it tobacco use versus smoking or is it whether the prohibition statement is effective *and this question is about the prohibition statement. It's not about whether its tobacco or smoking. So I'll chose 'yes' – the policy prohibits use on all school grounds.* (B, B8).

Excerpt 4. Uhh 'consideration' again. That's the same word that I saw last time that I didn't like…Uhhm…but this time it's stronger: it says 'out of consideration for the health and wellbeing of students, staff etc. *I'm gonna count that this time as intent…*(B, D8).

Excerpt 5. They don't say that it's prohibited but they say you receive a five day external suspension…*so I guess it's implied.* (B, F9).

Excerpt 6. …um, …looking for zero tolerance, (3 second pause) what happens if I smoke on school property. Says 'if you are sixteen or older you can be charged by Tobacco Enforcement Officer and be given a three hundred and five dollar ticket if you smoke on school property' (3 second pause) hmm so the Tobacco Enforcement Officer will.. give you a fine the first time they see you…but that doesn't mean that if you get caught by the Principal that they're gonna report you to the Tobacco Enforcement Officer... can I be, *oh there's another section. 'Can I be charged for just holding a friend's cigarette and not actually smoking?' and it says 'yes you will be charged if you are holding a lit cigarette on school property.'... So I will say yes to zero tolerance. If you are caught you will, you will be charged and consequences there are consistent* (B, E16).

Excerpt 7. …hm well it says that an Enforcement Officer will give you the ticket…huuh…but that's not really what we're thinking about. It's funny that they only talk about the consequences of breaking the act...instead of the consequences of breaking the school policy…..hmm…I'm wondering where the Principal fits into this, doesn't mention the Principal or any administrator from the school, or the school, hahaha all it talks about is the Act and the Enforcement Officers…and the Health Authority if you have questions, uhh who should be doing the disciplining? *Well since the only consequence is one that's given by the Enforcement Officer…then I guess they're the ones doing the disciplining…Says that 'you will be charged by Tobacco Enforcement Officers: I'll say yes.* (B, E17).

Excerpt 8....and s' I'm looking for how the, does it specify how the policy should be communicated to parents and guardians (three second pause) well it doesn't really specify how….uh mentions parental guardian permission…*no it doesn't say how. It doesn't say that there's signs or that they send a letter, if the students have to get it, permission from the parents...that doesn't mean that the parents become aware of the policy. There's more to the policy than just being allowed to smoke, in the designated area. Yeah.* (B, F7).

Excerpt 9.…so the intent of the policy is not, doesn't have to do with health but it has to do with safety…Yeah I'll check it out, well but do we want, do we want an intent that has to do with health, is that what we're haha asking for? Or just any intent...um…..we want an intent that has to do with smoking…..*'designated smoking area is to keep students who smoke with parental guardian permission safe…and better manage the students to keep them off of the highway and out of the woods. We do provide addictions counseling services at school and invite all students who require such services to please take advantage of them. We encourage all students to be* *non-smokers.'….. No I'm gonna say no to intent, doesn't seem to be the right kind of intent.* (B, F8).

Excerpt 10. (4 second pause). So much information makes me …- every time I look at another section I'm thinking back to the questions that I already answered haha whether I had noticed all of this but yeah, it all seems to be the same between the protocol and the board policy. It's just written differently…*so I'm just reading through it again…but there's nothing new popping out at me so there is no education and no cessation for students.* (B, L17).

Examining these key excerpts from both raters, it becomes clear that the main obstacles to rating are related to policy content and rating task content, as well as the link between the task and the policy content. Distinction of subtle details within policy content is the main obstacle in excerpt 1 for rater A. The rater takes a deliberate pause for contemplation of terms and decides to err on the side of caution by giving a conservative rating ('no'). Lack of clear information is another major obstacle related to policy content. The rater needs to answer a very specific question but often has only vague or limited information from a policy. Rater A overcomes this limitation by ruling out any chance of having overlooked this piece, acknowledging she does not have this information and assigning a 'no'.

A third type of rating obstacle occurs in relation to raters' memory. Occasionally, raters do not immediately remember a particular exception to the coding rules such as having to assign a score of -2 instead of zero or one. Raters overcome this type of obstacle by pausing to refresh their memory and re-access their insider knowledge in such a case.

The following section will present three cognitive process models of policy rating, based on the results for category use, rating obstacles and raters' coping strategies to deal with obstacles to rating.

Cognitive Process Models of Policy Rating

Figures 6 through 8 display cognitive process models of raters' policy rating responses; figure 6 shows the processes overall for both school-developed and board-developed policies, figure 7 shows the processes when policies rated are board – developed policies and figure 8 depicts the cognitive processes when policies rated are school – developed policies.

Figure 6. The cognitive process model of policy rating for all policies.

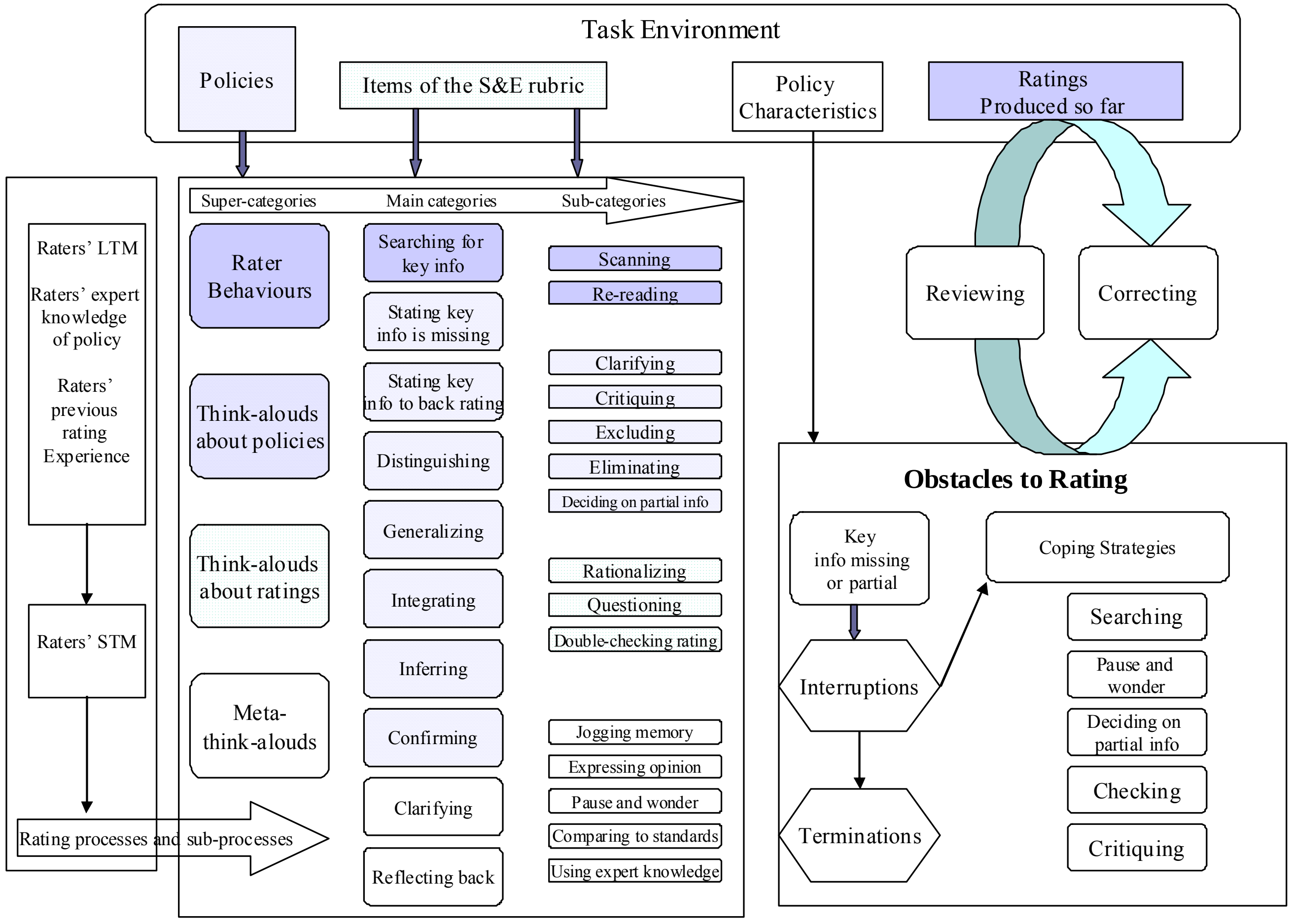


Figure 7. The cognitive process model of policy rating for board –developed policies.

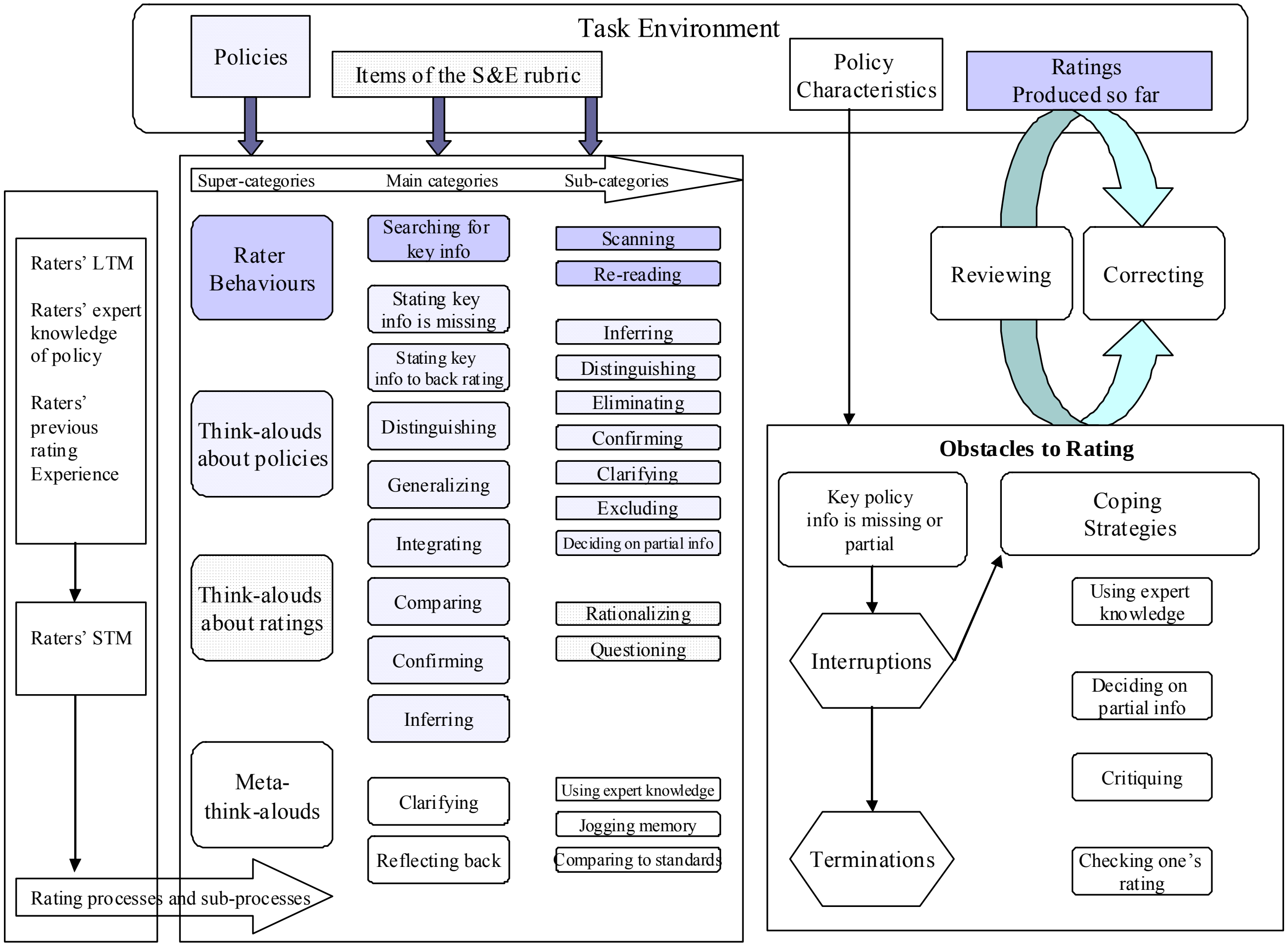
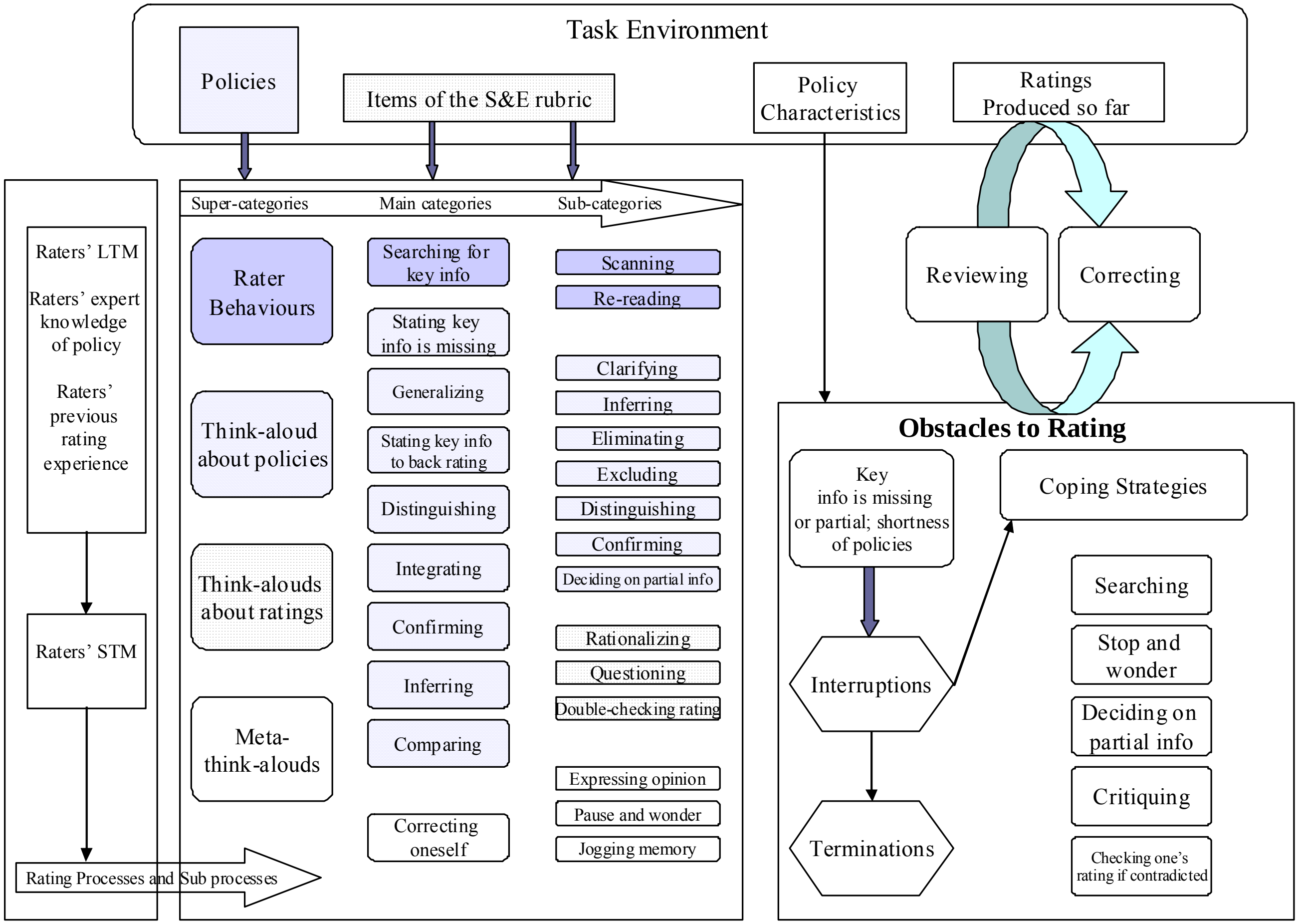


Figure 8. The cognitive process model of policy rating for school –developed policies.



Interpretation of the Cognitive Process Models.

Cognitive processes of policy rating in the cognitive models above were interpreted in the context of the task environment (top pane), which included the policies, the sample items of the Stephens & English (2002) rubric, policy characteristics as established in study one of this dissertation (Zeisser et al, 2009) and the ratings produced so far. Rating processes were also contextualized as influenced by the information processing conditions: raters' long term memory, their expert knowledge of the policies and their rating experience (left pane). However, one important point is that the information that can be verbalized has to be in the content of working memory (short-term memory). In other words, the content of the long-term memory (or the expert knowledge about tobacco policies) cannot be verbalized until it is retrieved. Hence, the left part of each model depicts this aspect of information processing and verbalization.

The middle part of each model shows the main rating processes and sub-processes, as established by the main-and subcategories extracted from the TA data. This level of detail was chosen to minimize the gap between the cognitive model and the protocol data. This part also depicts the four super categories these processes fall into. In essence, the super categories were based on the policies as a unit of analysis, the main categories were based on the items of the Stephens & English (2002) rubric as the unit of analysis, and the subcategories were based on finer-grained micro-units of thought within the item rating process as defined in the methods section. Hence, policies and items are the elements from the task environment that provide the direct linkage with the rating processes and sub-processes and thus, contextualize these processes. In addition, the hierarchy of the cognitive main-and subcategories reflects their frequency counts. That is, the categories observed with the highest frequencies appear on top of the hierarchy.

To further aid model interpretation in terms of category relationships, please note different textures of text boxes to emphasize category relationships and linkage with the task environment. Rater behaviours are highlighted in solid purple, TAs pertaining to the policies are stripe-patterned, TAs pertaining to ratings or items are dotted, and meta-TAs are left in white. Within each of these categories, the order of the main-and sub-processes reflects the frequencies with which these processes were observed in descending order. Hence, this structure shows not only the relationships between categories and which subcategories help better explain main categories; it also shows the context in which these processes are rooted and grounds categories in the data. That is, it can be glanced from the process models which sub-processes within what context best explain main categories in their context. For example, the model for all policies shows the category relationships between 'Rater behaviours', reflected in the main cognitive process of 'Searching for key information'. The subcategories to further explain the main category are 'Scanning the entire policy' and "Re-reading sections of the policy'. For further clarification of cognitive processes and their frequencies of occurrence, please also refer to Tables 6 to 8.

The second part of each model (right) shows obstacles that raters encounter and their coping processes in overcoming these obstacles. Policy characteristics and ratings produced so far are the other elements of the task environment that provide the direct linkage with rating obstacles and raters' coping strategies. That is, the processes of reviewing and correcting are used to show the interactions with the context of the task environment, in that the ratings produced so far are constantly being used by the raters in an iterative process. Policy characteristics are thus contextualized as interacting with the processes of dealing with obstacles to rating.

Variations in the rating processes as a function of policy type.

The models broken down by ratings of board-developed polices (Figure 7) and ratings of school-developed polices (Figure 8) showed subtle but important differences. That is, very similar rater behaviors were observed as main-and subcategories in all three models, but the TAs varied in their frequency, reflected in the category hierarchy. For meta-TAs, the model for board-developed policies shows the main processes ‘clarifying’ and ‘reflecting’ and the sub processes ‘using expert knowledge’, ‘jogging memory’ and ‘comparing to known standards’. These categories differ from the cognitive processes observed for ratings of school-developed polices alone. That is, the model for school-developed policies depicts the main process ‘correcting’ and the sub-process ‘double-checking’ for meta-TAs. These observed differences between the model for board-developed and school-developed policies are subtle but relevant in the subsequent explanation of differences in rater responding as function of policy type.

Discussion

The present study aimed to contribute to the discourse about validity of inferences from rating scores by focusing on the process of rater responding. Specifically, the study addressed the substantive aspect of construct validity (Messick, 1995). Messick's notion of the substantive aspect of construct validity and his call for the study of process models were a particularly useful foundation for the context of policy ratings in that a series of cognitive process models were proposed to enhance understanding of rating score meaning by illuminating process components of expert rater knowledge. The direction taken by the present study is also in keeping with the Standards of Educational and Psychological Testing (AERA, APA, NCME, 1999), which call for the study of cognitive processes of responding as an important approach to gathering evidence for the validity of score inference and meaning. Careful study of these cognitive process models enables researchers to arrive at decisions regarding the validity of inferences based on scores given by the expert raters. Specifically, the cognitive process models of rating developed in the present study via the Stephens & English (2002) rubric help inform an evaluative judgment regarding the usefulness of this measure in school tobacco policy rating and the practical implications for those who apply this measure in the quest to quantify policy strength and impact on health outcomes.

Examination of the process models revealed that the cognitive processes and strategies used by expert raters to rate written tobacco policies were of a wide variety. These processes were used with great range of frequency and included assessing information completeness, formulating questions and distinctions, and generating inferences. Hence, these strategies are generally comparable to the strategies found by researchers of other cognitive processes, such as information processing and reading comprehension (Ericsson & Simon, 1993; Olson, Duffy &

Mack, 1984; Pressley & Afferbach, 1995). This section will discuss the raters' use of process categories and the relationship of super-, main- and subcategories in the context of the rating act (e.g., task environment and policy characteristics) and with an eye towards adding to the validity argument pertaining to inferences made from these ratings. These results are discussed in the context of the cognitive models shown in Figures 6 to 8. Further, the results from the present study are discussed in relation to the tobacco policy characterization study of this dissertation (Zeisser et al., 2009), since it provided the conceptual basis for the present study. It is emphasized that the combinations of rating items and rating objects (e.g., tobacco policies of varying types, lengths and comprehensiveness) are a strong foundation for the validity arguments formulated in the present study and formed major components of the cognitive models developed. In addition, it is important to note that the policies sampled for the TA study were drawn from a randomized stratified sample within the context of the 2006/07 Youth Smoking Survey. Hence, the policies can be viewed as representative of the Canadian school tobacco control policies. The following part highlights the importance of process model interpretation in light of policy characteristics.

Variations in the rating processes as a function of policy type.

The comparison of the overall model of rating with the models for school-developed policies and board-developed policies showed only minor differences and the overall model serves more or less as a general description of tobacco policy rating. However, the analyses that differentiated by policy type showed that there are interesting variations in the processes of rating school-developed policies, as opposed to board-developed polices and these variations speak to the rating objects. While very similar rater behaviors were observed as main-and subcategories, the TAs varied in their frequency. That the main process of 'comparing to known standards' occurred more frequently when board-developed polices were rated can be understood as an interaction between rating and the nature of the object being rated. That is, board policies – being generally longer and more comprehensive – contain more statements that allow such comparisons. That this main process occurred with less frequency in the model for school-developed policies is an indication that raters are less able to make such comparisons due to lack of information. This finding is also in line with the observation that raters' stating that key information needed to positively answer the item (with 'yes') is missing occurred 41 times with school-developed policies but only 31 times with board-developed polices. It was also an expected finding that the main process of 'generalizing' occurred more frequently for rating school-developed policies. School-developed policies tend to be much shorter and less comprehensive and thus, raters are more likely to use generalizations to come up with a rating response. However, for the same reason, it was surprising that the main process of 'inferring' occurred with almost the same frequencies for both policy types. Conversely, the process of 'integrating multiple pieces of policy information' was observed more often in the board-policy scenario, where raters frequently had to process a large amount of information available due to the greater comprehensiveness of this type of policy compared to school-developed policies.

The difference with respect to TAs pertaining to the ratings can likewise be explained by contextual variation in the task environment. For example, when rating school-developed policies, raters used the sub-process of 'double checking one's ratings', an indication of uncertainty about the rating decision made earlier. For board-developed polices, raters never used 'double checking'; this is a subtle difference, but it shows that raters experience differing degrees of comfort with their decision process depending on the rating context – the characteristics of the object being rated (Figures7 and 8).

For meta-TAs, the main processes unique to rating board-developed policies – 'clarifying terms and meanings' and 'reflecting back on one's understanding of policy content' can also be interpreted as a contextual difference in that raters' thinking is stimulated more when longer and more comprehensive policy content is available. In other words, the absence of these main processes when experts rate school-developed policies can be explained by the design difference and the resulting change of context: when the objects rated are school-developed policies, certain cognitive processes are not used by raters due to the characteristics of the policies. A similar explanation can be formulated based on a cognitive main category unique to rating school-developed policies: 'correcting oneself if contradicted' occurs only during rating policies with certain characteristics, and one explanation could be that contradicting information is more likely to be found in the shorter, less comprehensive school policies. With respect to the meta-TAs, it was an expected result that raters more frequently 'stopped and wondered' or expressed a personal opinion when rating school-developed policies, but relied more on their expert knowledge when rating board-developed policies. When expressing a personal opinion, raters frequently referred to a lack of detail in information needed for rating school-developed polices, or complete absence of such information. This was also found in the analysis of excerpts for rating obstacles and rater coping to overcome these. A rather unexpected result was the use of the main process of 'clarifying terms and meanings' when rating board-developed policies and the absence of this process in the model for rating school-developed policies. This speaks to the observation that even though more comprehensive information is available to raters when rating board-developed policies, this information is not necessarily clear and succinct, so that raters need to employ a process of clarification before they move on to assigning a rating.

Rating obstacles and raters' coping strategies.

As with the main-and sub-processes of rating, the context of what was being rated also provided clues with respect to rating obstacles and how raters coped with these. While the main key obstacle to rating either type of policy was that key information needed was partial or missing, the difference of rater coping speaks to the importance of the rating context. While raters frequently critiqued both school-and board developed polices for their lack of clarity or information and hence, had to fall back on deciding on partial information, they where able to frequently apply their expert knowledge when rating board-developed policies, but did not use this process at all in rating school-developed polices. Rather, for the latter, raters predominantly spent their efforts searching the policy for more information on which to base the rating. This result indicates that board-developed policies, which were generally longer and more comprehensive, contained sufficient information so that raters could overcome obstacles by using their expert knowledge to fill in the gaps. For school-developed policies, which were generally shorter and less comprehensive, raters were not able to fall back on their expert knowledge to bridge the lack of information provided in the policy. Instead, raters focused their coping efforts on searching the policy once again for the information specified in the rating item to be certain to not have missed any content before assigning a 'no' (0) rating.

Another interesting finding was the use of the process 'checking one's rating if contradicted' when rating school-developed policies – a process that did not occur when rating board-developed polices. While this process only occurred four times, it nevertheless speaks to the interaction between the object being rated, the rating context and raters' decision processes. This result indicates that the validity argument built around inferences from rating must take into consideration the elements of objects being rated, items, rating context and raters themselves. In summary, the finding of differences in cognitive processes when rating school-developed or board-developed polices is important since it indicates that raters indeed engage in cognitive processes reflective of domain content. Hence, this finding speaks to the substantive aspect of construct validity (Messick, 1994, 1995), highlighting the identification of domain processes revealed in the rating task.

What can one learn from these results on rating obstacles and raters' coping strategies? With respect to rating obstacles, one also needs to consider the role of working memory. Problems with working memory and synchronization of verbalizations can be recognized by interrupted verbalizations. Hence, for the validity of inferences from ratings one needs to acknowledge that some of the expert knowledge may not be reflected in the scores due to less-than optimal flow between working memory content and verbalizations. It is strongly emphasized here that it was not the aim of this study to criticize the quality of the objects being rated – the policies, or critique either type of policy. Rather, the separate content analyses by type of policy were useful in showing and understanding the different cognitive processes in raters when rating different polices.

Another point is strongly emphasized and pertains to the relationship between the present TA study and the policy characterization study of this dissertation (Zeisser et al., 2009). The findings showed various interactions between policies and Stephens & English (2002) items in cognitive category formation. This result highlights that if one approaches the substantive aspect of construct validity via the process of responding, the elements of the task environment as they were established in the policy characterization are of crucial importance for score meaning and process model interpretation, as is the raters' interaction with these elements. For example, notice the cognitive feedback loop that forms between obstacles to ratings and another element in the task environment – ratings produced so far. This cognitive feedback loop is used extensively by both raters to deal with rating dilemmas during the process of responding. This is very important since these rating dilemmas are resolved in one of two ways. The first way is via interruptions. These interruptions lead back to raters' coping strategies via cognitive process categories such as searching, checking, critiquing etc., whereby the ratings produced so far are utilized, as well as the polices already rated. The second way in which coping dilemmas are resolved is via terminations of the rating process by assigning a 'no' rating when the above described coping strategies yield no result. In summary, the findings and the proposed cognitive models of rating clearly show the importance of considering all the elements of the task environment, as well as interactions of raters with these, when formulating the validity argument via the process of responding. These points also highlight the importance of the groundwork that forms the basis of such a validity approach for the process of rating in particular. This groundwork involves careful study and characterization of rating objects, selection of rating items and the entire set of elements comprising the task environment. It is for this purpose that the policy characterization study of this dissertation was specifically important. The following section will discuss some alternative explanations as well as the present study's strengths and limitations.

Alternative explanation, limitations and strengths of the study.

This study has several limitations. First, only two expert raters were available to rate Canadian tobacco policies. Clearly, having a larger sample of raters would increase the value and strength of the cognitive process models and hence, any inferences one could make from the policy ratings about rating score validity. Since there were only these two expert raters available, the raters had already seen and rated the sample of policies in the present study approximately four months earlier in the context of the YSS. While highly unlikely, the possibility of memory and recency effects is acknowledged in that having rated these policies before, the TA results may have been influenced by the raters' prior rating experience.

Another limitation generally associated with TA studies is the possibility that raters try to give a desirable response. That is, there is no guarantee that the verbalizations accurately reflect raters' thought processes at the time. In line with the general tone in the literature (e.g., Young, 2005), it is acknowledged that during the TA process, when a rater is absorbed in a given activity, the completion of this task will take precedence. This could potentially limit the available cognitive resources for the rater.

A related limitation is that the task of rating the sample of tobacco policies for validation purposes may have caused a performance-oriented situation for the expert raters in that they may have felt some pressure about their knowledge of policy and the rating process, even though both raters were previously assured that this task was not a job performance - related task. Perceived pressure of this nature could result in raters rating differently than in their usual job situation.

Transcripts are complex, and raters' statements can be ambiguous or incomplete when they describe what they are doing. As a result, there is the potential for having to make judgments about the raters' meaning. If such judgments occur, this may pose a threat to the objectivity of the analyses. It is acknowledged that thought proceeds much faster than speech, and that TAs are inherently incomplete renditions of thought. In addition, no causal relations can be inferred between constructs from the content analysis. That is, even though the categories were grounded in the data, the frequency counts cannot and should not be used to draw causal inferences.

Finally, an important but rarely addressed issue in TA applications and subsequent protocol analyses is the role of the researcher as part of the interpretation and analysis process. That is, the researcher may view the results through a lens influenced by his or her background and theoretical orientation. It is acknowledged that the present study did not contain a specific analysis of this potential influence and that results may reflect, at least to a certain extent, the interpretive lens of the researcher.

This study also has various strengths. Rather than having a lot of 'breadth' with respect to the number of raters, the study focused on depth by ensuring that a sufficiently large and comprehensive sample of written policy material was rated. Further, both raters were trained and experts. It is emphasized that the interest was in understanding and modelling the cognitive processes of expert policy raters, rather than people who are learning to rate. Hence, there was a trade-off between breadth and depth in order to be able to benefit from access to expert knowledge in policy rating.

In addition, the study was based on a rigorous policy sampling developed in the policy characterization study (Zeisser et al., 2009), and implemented to ensure that all types and facet combinations of policy documents naturally occurring are rated. Further, an item sampling frame was implemented so that the content domain of the Stephens & English (2002) rubric was covered broadly, as well as ensuring coverage of the items that are the most difficult for raters. A rigorous methodological procedure was followed to ensure transparency throughout the entire study. For this purpose, strict methodological guidelines previously developed by Tong et al. (2007) were followed to make every attempt to avoid flaws and errors. A reliability check was conducted by a second independent coder to verify the categories developed by the first author.

Discrepancies and gaps were identified, discussed in detail and addressed before analyses proceeded.

Future research.

Next steps for research should focus on replicating and testing the validity of cognitive process models of ratings. The degree to which score meaning and action implications hold across respondents and populations as well as across settings and contexts is a persistent empirical issue for researchers. Hence, the cognitive process models presented here for different types of policies should be replicated with more trained raters and more policies. The hope is that researchers will be inspired to include an examination of qualitative TA data in their explorations of policies, either through content analysis of TA protocols developing cognitive process models of rater responding or expert interviews, to more completely capture the meaning of raters' decisions. Future researchers are also encouraged to consider a variety of coding schemes at either a more micro or more macro level than presented here, for examining participant-generated justifications for ratings and decisions. A continued research focus on the thought processes of raters that contribute to policy-related decisions can advance our understanding of the validity and fairness of these decisions. Such a focus would further add to the explanation-oriented view of validity and validation (Zumbo, 2009), as it would enhance understanding of the processes involved in rating score generation.

Implications for practice and recommendations.

Despite the limitations of this study, there are some notable implications that speak to the use of the Stephens & English (2002) rubric for rating tobacco policies. First, one can conclude that this measure worked well for policies that were long and comprehensive in detail (generally the board-developed policies). Here, it becomes immediately apparent how important it was to characterize tobacco policies in study one of this dissertation (Zeisser et al., 2009), where key policy characteristics (type, length and comprehensiveness) were identified with the aim to lay out and define the task environment for the TA protocol. The Stephens & English (2002) coding rubric is a very detailed tool and addresses five major policy areas (e.g., how the policy was developed, tobacco-free environments) with a variety of rating items that address specific groups (e.g., students, teachers, visitors). When tobacco policies are of the long and comprehensive type, raters tend to be in possession of the information they need in order to rate policy content with the Stephens & English (2002) rubric items; raters can easily discern whether or not the policy met the criterion specified by the rating item.

However, in many cases (generally with school-developed policies), raters are faced with less-than ideal school tobacco policy documents with respect to length and comprehensiveness of information provided. That is, currently, some Canadian school tobacco policies are not yet in the form that would allow raters to make full use of the Stephens & English (2002) rubric with respect to capturing the maximum possible content information about policy strength. This is an important finding since this issue concerns measurement quality. In such task conditions, raters are not able to provide much TA data and generally just verbalize that the information is not mentioned in the policy. Raters are then forced to give a 'no' (0) rating on this item. In these cases, the Stephens & English (2002) rubric still works well as a measure; a short policy lacking comprehensiveness would receive a low score due to raters not being able to give a rating of yes (1) on the majority of rating items because the required information is not available. It is strongly emphasized that the above issue of raters having to give a 'no' rating is not a flaw of the Stephens & English rubric, but rather a reflection of policy content, specifically the lack of detail provided in the policy. That is, since the relevant content described in the Stephens & English rubric items is not mentioned in the policy (e.g., “Does the policy prohibit smokeless tobacco”?), raters can only assign a ‘no’ rating. The implication for practice is that the Stephens & English rubric still works well as a measurement tool as such when policies are short and have little detail; the rubric assigns a low total score for the policy.

In light of the findings, one could conclude that schools need to develop longer, more comprehensive policies in the future, with the scope of the rubric in mind. However, one needs to weigh such a recommendation carefully with potential positive *and* negative consequences of assessment in mind (Messick, 1995). While one can create longer policy documents, one could do so without capturing the specific content that is necessary to achieve reduced smoking rates. In addition, a potential unintended negative consequence could be that measurement, on its own, would have too heavy a hand in policy development, leading to the proverbial ‘tail wagging the dog’. That is, while a potential positive consequence is that schools indeed develop longer and more comprehensive tobacco policies, a potential negative consequence is that schools would be developing policies simply in a ‘teaching to the test’ fashion. In addition, while policy intent may be strong on paper, effective policy implementation and enforcement are required to accomplish lower smoking rates.

What is noted though is that there exists a discrepancy between the rubric’s detailed scope and the actual information provided in the rating objects. That is, a ‘measurement gap’ was found between the level of detail in the measure and the level of detail in some measurement objects (policies). Hence, it is theoretically possible that a school tobacco policy receives a low total score due to lack or absence of key information about smoking as specified in the rubric. For tobacco policy researchers, this is a concern since such a score may contain ‘error’ by lack of information; the score is as weak as, say the score of a policy that clearly states ‘our school provides smoke pits for students and teachers'. However, the school with the low-scoring policy due to lack of information may in reality have smoke-free environments, prevention and cessation programs and hence provide a better environment for students; we just cannot know this because the written policy document does not contain the information and detail. There can only be very limited understanding of what such a score means and hence, explanation is also limited. It follows that inferences from such policy scores (e.g., about smoking outcomes) are very limited and potentially inappropriate.

It follows that in the case of long and comprehensive school tobacco policies, the cognitive models better support score interpretation because these scores have meaning: there is sufficient congruency of rating object and rating tool content domain coverage. Hence, we are able to see a variety of cognitive process components that lead to the score. Through this, we obtain a better sense of raters' thinking by focusing on the cognitive process itself. The processes are meaningful because they relate to the objects rated – the policies - and because they help explain rater responses.

Conversely, in the case of short policies with minimal comprehensiveness, there also tends to be limited output of cognitive processes, for example when raters terminate the process by assigning a score of zero since they cannot find the information in the policy. Alternatively, raters may engage in extensive coping strategies and verbalize these out aloud. This information was useful in the generation of the process models; such information provides clues and insights about variations in rating scores by illuminating how these variations arose in the process of rating. This adds to the substantive aspect of construct validity as outlined by Messick (1994, 1995), and to the explanation-oriented view of validity emphasized by Zumbo (2009).

What is the implication for other researchers regarding the use of the Stephens & English coding rubric? Findings of different cognitive processes in expert raters when rating school- versus board-developed tobacco policies has highlighted two main points with respect to use of the rubric. First, this result demonstrated that raters indeed engaged in cognitive processes reflective of policy domain content available to them. Second, the lack of depth regarding TA responses when rating school-developed policies (e.g., giving predominantly brief responses indicating that the needed information to assign a 'yes' rating was absent) *cannot* be interpreted as a limitation of the assessment tool – the Stephens & English rubric. Instead, this result is due to actual policy domain content provided in the school-developed policy documents – the objects of rating. Therefore, the inferences one can draw from policy scores based on the Stephens & English coding rubric as a tool for measuring school tobacco policy strength are strengthened as justified, accurate and appropriate. Nevertheless, more insights into the process of rater responding can be gleaned with respect to rating board-developed policies specifically, since the TA responses from this type of policy were richer and contained more information about how raters went about assigning the scores. One can attribute this richness in TA responses to the larger scope of relevant content provided in board-developed policies in general, in that it stimulated more engagement in thought processes reflective of domain content during the assessment task. Evidence of participants' engagement in cognitive processes reflective of domain content is a key element speaking to the substantive aspect of construct validity; it shows content representativeness of the construct measure and process representation of the construct (Messick, 1995). In other words, these response consistencies show the degree to which the processes are reflected in construct measurement.

With respect to the use of the Stephens & English rubric and the validity of inferences from policy rating scores, the present study also touched on the structural aspect of construct validity (Messick, 1995). That is, the TA study results showed that the scoring model was reflective of task and domain structure. In other words, the scoring rubric was shown to be rationally consistent with the construct domain of policy quality and to contain construct-based scoring criteria.

The findings and conclusions also have practical implications for tobacco policy development. By illuminating raters' thought processes, the present TA study also told the story of which are the better school tobacco policies in showing how raters think while rating. It is recommended that researchers wishing to quantify policies collaborate with officials developing school tobacco polices (e.g., school advisors, principals, school boards and school health advisors). The focus needs to be on developing school tobacco policies that are long and comprehensive enough to clearly specify key policy aspects such as smoking prohibition, locations of prohibition, tobacco prevention education and access to smoking cessation programs. As Stephens & English (2002) state, a strong school tobacco policy must provide this extent and level of detail. However, excellent policy intent is not sufficient – one also needs to consider the importance of policy implementation. For research practices, this means that if school tobacco control policies are to be quantified and scores used in predicting smoking outcomes, the characteristics of such policy documents need to be taken into consideration; they need to be addressed at the time when policies are developed. For this purpose, it would be advisable to identify the strongest, most ideal school tobacco control policies for use as guidelines for others to follow when developing their policies. Specifically, policy makers need to develop policies that provide sufficient detail and hence, would be more measurable using the

Stephens & English (2002) coding rubric. For this specific purpose of improving school tobacco policies, it is recommended that future tobacco policy studies include an element of liaison between those who quantify policies with the aim of predicting smoking outcomes and those who develop them (e.g., school boards, administrators or principals). Currently, there appears to be no guideline for school tobacco policy developers with respect to basic minimum criteria – aside from the actual Stephens and English (2002) coding rubric itself. Hence, it is also recommended that the rubric be used as a guideline; to this end, the rubric needs to be disseminated to school tobacco policy developers. Such a guideline would be helpful in creating a set of school tobacco policies that would allow for better quantification and more fair comparisons. This could lead to better practice regarding Canadian school tobacco policy overall. If one has knowledge of what the best and strongest school tobacco policy is, one can develop and implement policies accordingly, reduce smoking rates at the school age in the short term and in the long term, hopefully reduce the burden of disease. Further, an improved tobacco policy system for Canadian schools would benefit researchers interested in exploring the relationship between school tobacco policies and youth smoking outcomes. In this endeavour, one also needs to distinguish clearly between policy intent and policy implementation. That is, while it is possible to develop comprehensive policies with excellent intent (e.g., to provide healthy smoke-free environments to students and teachers), this does not ensure that the policy is implemented well enough to produce the desired outcomes.

In addition, a positive consequence is the use of the Stephens & English rubric in context and based on how this rubric was developed. That is, the rubric was developed using health education and promotion as a standard for content. Therefore, the Stephens & English rubric can become a standard on what we consider a strong tobacco policy. In light of Messick’s structural aspect of construct validity, this would be a positive consequence because the rubric's development has best practice of current health promotion as a foundation.

The findings of the TA study have potential implications beyond tobacco prevention; they are of relevance for rating of policy in general. Examples of other areas would be the rating of environmental, social or health policies. The area of tobacco research served as a good example how the method of TA can be applied to illuminate rater processes in other fields. This would be useful since it would create a vibrant research environment and knowledge exchange across disciplines. In addition, policy characteristics such as scope and comprehensiveness can now become variables in policy evaluation studies and policy research in general. The policy characteristics examined in the present TA study represent but a small set of many variables that could serve as policy characteristics in other fields.

Finally, the present study has implications for rater training. On a broad level, the importance of policy characteristics for producing different ratings and decisions can help inform policy researchers about what their raters are considering and focusing on during the task. This information can then be utilized in future rater training, where raters could be alerted to commonly found tendencies to pay attention to particular characteristics when rating polices of various types in order to reduce unfair bias towards either school-developed or board-developed polices. Raters could also be trained to cope with commonly found rating obstacles using the information from the present study.

Implications for rater training also should be seen in light of Messick's (1990) view of validity and his focus on potential social consequences of score use. That is, the examination of the consequential basis for rating score use requires an appraisal of both the potential and actual social consequences of score use and a consideration of value implications. In addition to evidence supporting construct validity and value implications, consideration also needs to be given to relevance and utility of the rating scores. For example, future rater training can be designed with the utility, value implications and social consequences for students, schools and policy developers in mind.

Lastly, with respect to future rater training, it became clear that the expertise and extensive training that the raters in this study had helped them to ‘fill in the gaps’ and cope with the challenge that many policies lacked information they needed to rate, or had only partial information. Hence, the use of trained raters is strongly recommended when the rating objects are challenging documents such as tobacco polices that are often short in scope and limited in detail and comprehensiveness.

To summarize, I recommend using the Stephens and English rubric in school tobacco policy coding, since the tool worked for long and comprehensive school policies by eliciting more “yes” responses; it also worked well for short policies lacking comprehensiveness, by eliciting more ‘no’ responses from the raters. However, caution is necessary in score interpretation and use. That is, there needs to be sufficient correspondence between the level of the measurement tool (content domain coverage) and the level of information contained within the rating objects themselves. As Messick (1994) emphasizes, the substantive aspect of construct validity stresses two important points: the need for tasks providing appropriate sampling of domain processes and the need to move beyond traditional judgment of content to amount empirical evidence that respondents indeed engage in the sampled processes. The validity of inferences from policy scores can only be as strong as the information gleaned from cognitive processes of responding. Whenever raters had to respond with “the information I need to rate this component is not mentioned”, one needs to consider this missing input into the process model of rating. The models proposed in the present study are hence also limited to the degree that only so many rater responses were available to draw out the process components; other components could have arisen with more information available. The following section discusses the novel contributions of this study.

Novel Contributions of the Think-Aloud Study.

Results from the TA study make novel contributions to the fields of assessment and tobacco policy research in several different ways. First, the results illuminate some of the cognitive processes underlying the responses of raters when they code tobacco polices using the Stephens & English (2002) coding scheme. The use of the TA method for illuminating expert rater knowledge is an important contribution to the discussion about the validity of inferences from rating scores. Insights about how these experts perform their task inform researchers working with the scores with respect to what can be said about the nature of these scores. Specifically, the findings from the TA study on cognitive main-and sub-processes allowed for the development of three separate cognitive models showing the process of tobacco policy rating. Hence, a deeper understanding of the response processes in generating tobacco policy rating was gained. This information allows for the gathering of validity evidence via the explanation of rating variations. The TA method was applied in a new way as a means to gather validity evidence and build a foundation for constructing theories of rating processes. Hence, the findings and proposed models could form a basis from where to develop cognitive process theories of policy rating.

The second novel contribution of the TA study is that results about the cognitive categories employed by the tobacco policy raters contribute to gathering validity evidence to support inferences made from the policy rating process. This is accomplished by relating the cognitive processes of responding to the objects rated – written tobacco policies - , and their various facets in a systematic way, thereby improving the measurement quality (Zumbo, 2005) of the tobacco policy rating process. This issue has not been studied before, but should be researched according to Cizek et al. (2008), who warn that only a very small fraction of social science reports make use of the cognitive processes in respondents to better understand the answers they produced. The practical contribution of these results to assessment is emphasized when constructive input into policy recommendations is made by drawing on raters' cognitive processes. Greater understanding of how the tobacco policy ratings were generated inspires greater trust in those ratings and hence, the ratings provide a more useful input to those who aim to improve pragmatic policy decision making.

The third novel contribution is that results from this study are also expected to enhance the application of the quantitative policy indicators constructed from the ratings. The intended future use for these indicators is for them to be part of a readily accessible tool for schools to evaluate their own smoking polices (future work). Along these lines, the cognitive models of rater responding can also be used as a basis to form new hypotheses and research questions in the area of quantitative policy assessment in tobacco research.

The fourth novel contribution of this study pertains to future rater training in tobacco policy rating contexts specifically, but also in other rating contexts more generally. Many of the cognitive processes discovered through this TA study are teachable in settings other than tobacco research. Therefore, the results can potentially help inform how future rater training is conducted and what type of training protocol will be developed, for use in tobacco policy- or other rating contexts.

The fifth novel contribution is that the methodology section describes in great detail a novel approach to construct validation and the substantive aspect of construct validity via the process of responding; since this methodology is rarely used in construct validation, there are no models or guidelines for researchers to follow. The detailed description may serve other researchers as a starting point or methodological guideline for future studies.

Finally, the TA study makes a contribution to interdisciplinary research by applying methods used traditionally in education and psychology to a research project at the intersection of applied population health sciences, statistics and psychometrics.

References

Aanstoos, C. M. (1983). The think aloud method in descriptive research. *Journal of Phenomenological Psychology, 1*4(2), 243-266.

Adams, M. L., Jason, L. A., Pokorny, S., & Hunt, Y. (2009). The relationship between school policies and youth tobacco use. *Journal of School Health, 79*(1), 17-23.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). The Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Anderson, R. C. & Pearson, P. D. (1984). *A schema-theoretic view of basic processes in reading comprehension.* Bolt, Beranek & Newman, Inc., Cambridge, M.A.

Appleton, J. V. (1995). Analyzing qualitative interview data: Addressing issues of validity and reliability. *Journal of Advanced Nursing, 22*(5), 993-997.

Backlund, L., Skånér, Y., Montgomery, H., Bring, J., & Strender, L. (2003). Doctors' decision processes in a drug-prescription task: The validity of rating scales and think-aloud reports. *Organizational Behavior and Human Decision Processes, 91*(1), 108-117.

Baker, L., & Brown, A.L. (Ed.). (1984). Metacognitive skills and reading. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 353-394). White Plains, NY: Longman.

Banning, M. (2008). The think aloud approach as an educational tool to develop and assess clinical reasoning in undergraduate students. *Nurse Education Today, 28*(1), 8-14.

Bartolone, J. (2004). Using think-aloud methods to understand text comprehension. US: ProQuest Information & Learning.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1-17.

Beach, R., & S. Hynds. (1991). Research on response to literature. In R. Barr, M. Kamil, P. Mosenthal & P. D. Pearson, *Handbook of reading research*, Vol. II , New York, N. Y.: Longman.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*(1), 77-92.

Berg, B. L. (2007). *Qualitative research methods for the social sciences.* (6th ed.). Boston, MA: Pearson International.

Berne, J. (2004). Think-aloud protocol and adult learners. *Adult Basic Education, 14*(3), 153-173.

Block, C. C., & Israel, S. E. (2004). The ABCs of performing highly effective think-alouds. *Reading Teacher, 58*(2), 154-167.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an advanced placement history examination. *Journal of Educational Measurement, 31*(4), 275-293.

Cannella, G. S. (1992). Gender composition and conflict in dyadic interactions: Effects on spatial learning in young children. *Journal of Experimental Education, 61*, 29-41.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal, 26*(4), 499-531.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Coffman, W. E. (Ed.). (1971). Essay examinations. In R. L. Thorndike (Ed.) *Educational measurement* (2nd ed.). Washington, DC: American Counsil on Education.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Daly, K. J. (2007). *Qualitative methods for family studies & human development.* (1st ed.). Thousand Oaks, CA: Sage Publications.

Davey, B. (1983). Think aloud-modeling the cognitive processes or reading comprehension. *Journal of Reading, 27*(1), 44.

De Beaugrande, R. (1981). Design criteria for process models of reading. *Reading Research Quarterly, 16*(2), 261-315.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.

Dewey, J. (1910). *Ho we think.* Boston: Heath.

Dey, I. (1993). *Qualitative data analysis: A user-friendly guide for social scientists* (1st ed.). New York, NY: Routledge.

Dunker, K. A. (1926). A qualitative study on productive thinking. *Journal of Genetic Psychology, 33*, 642-708.

Elstein, A. S., Schulman, L. S. & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. & Gorin, J. (2001). Improving construct validity with cognitive psychological principles. *Journal of Educational Measurement, 38*(4), 343-368.

Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology, 16*(8), 981-987.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports on thinking. In C. Faerch, G. Kasper, C. Faerch & G. Kasper (Eds.), *Introspection in second language research*. (pp. 24-53). Clevedon England: Multilingual Matters.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: The MIT Press.

Evans-Whipp, T., Beyers, J. M., Lloyd, S., Lafazia, A. N., Toumbourou, J. W., Arthur, M. W., et al. (2004). A review of school drug policies and their impact on youth substance use. *Health Promotion International, 19*(2), 227-234.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative Health Research, 3*(4), 430-441.

French, D. P., Cooke, R., McLean, N., Williams, M., & Sutton, S. (2007). What do people think about when they answer theory of planned behaviour questionnaires? A 'think aloud' study. *Journal of Health Psychology,12*(4), 672-687.

Fuchs, L. S., Fuchs, D., Bentz, J., Phillips, N. B., & Hamlett, C. L. (1994). The nature of student interactions during peer tutoring with and without prior training and experience. *American Educational Research Journal, 31*(1), 75-103.

Funkesson, K. H., Anbäcken, E., & Ek, A. (2007). Nurses' reasoning process during care planning taking pressure ulcer prevention as an example. A think-aloud study. *International Journal of Nursing Studies, 44*(7), 1109-1119.

Gay, L. R. & Gallagher, P. D. (1976). The comparative effectiveness of tests versus written exercises. *Journal of Educational Research, 70*, 59-61.

Ghaith, G., & Obeid, H. (2004). Effect of think-alouds on literal and higher-order reading comprehension. *Educational Research Quarterly, 27*(3), 49.

Glaser, B. G. (1965). The Constant Comparative Method of Qualitative Analysis. *Social Problems, 12*(4), pp. 436-445

Harwell, M. (1999). Evaluating the validity of educational rating data. Educational and *Psychological Measurement, 59*(1), 25-37.

Hayes, J. R. & Flower, L. S. (1983). Uncovering cognitive processes in writing: An introduction to protocol analysis. In P. Rosenthal, L. Tamor & S. A. Walmsley (Eds) *Research on writing: Principles and methods*.. New York: Longman.

Hayes, J. R., Flower, L., Schriver, K. A., Stratman, J. F., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, vol. 1: Disorders of first-language development; vol. 2: Reading, writing, and language learning.* (pp. 176-240). New York, NY US: Cambridge University Press.

Heerkens, H., & Van Der Heijden, B. (2005). On a tool for analyzing cognitive processes using exploratory think-aloud experiments. *International Journal of Human Resource Development & Management, 5*(3), 1-1.

Hogan, T. P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802-812.

Holsti, O. R. (1968). Content analysis. In G. Lindzey & E. Aaronson (Eds.), *The handbook of social psychology*. Reading, MA: Addison-Wesley.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Jaspers, M. W. M., Steen, T., Bos, C. v. d., & Geenen, M. (2004). The think aloud method: A guide to user interface design. *International Journal of Medical Informatics, 73*(11), 781- 795.

Joseph, G., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making, 10*(1), 31-44.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4, Measurement Update for the 21st Century), 319-342.

Kane, M. (Ed.). (2006). Validation. In R. Brennan (Ed.) *Educational measurement* (4th ed.). Washington, D.C.: American Council on Education.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29*(2), 303-323.

Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: Analysis of a protocol. *Cognitive Science, 8*(4), 363-385.

Kuipers, B., Moskowitz, A. J., & Kassirer, J. P. (1988). Critical decisions under uncertainty: Representation and structure. *Cognitive Science: A Multidisciplinary Journal, 12*(2), 177-210.

Lewins, A., Taylor, C. & Gibbs, G. R. What is qualitative data analysis? Accessed online July 2, 2009.

Lovato, C. Y., Sabiston, C. M., Hadd, V., Nykiforuk, C. I. J. & Campbell, S. H. (2007). The impact of school smoking policies and student perceptions of enforcement on school smoking prevalence and location of smoking. *Health Education Research, 22(6)*, 782-793.

Lucas, E. J., & Ball, L. J. (2005). Think-aloud protocols and the selection task: Evidence for relevance effects and rationalization processes. *Thinking & Reasoning, 11*(1), 35-66.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.

Maykut, P. & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide* (1st ed.). Washington, DC: The Falmer Press.

Messick, S. (Ed.). (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.). New York, N. Y.: Macmillan.

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. (Research Report No. RR-94-45). Educational Testing Services, Princeton, NJ.

Messick, S. (1995). *Standards of validity and validity of standards in performance assessment*. Educational Testing Services, Princeton, NJ.

Meyers, J., Lytle, S., Palladino, D., & Devenpeck, G. (1990). Think-aloud protocol analysis: An investigation of reading comprehension strategies in fourth- and fifth-grade students. *Journal of Psycho-educational Assessment, 8*(2), 112-127.

Miller, G. A. (1962). Psychology: *The science of mental life.* New York: Harper & Row.

Miller, S. I. & Fredericks, M. (1996). *Qualitative research methods: Social epistemology and practical inquiry* (2nd ed.). New York: NY: Peter Lang Publishing.

Montgomery, H., & Svenson, O. (1989). A think-aloud study of dominance structuring in decision processes. In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making.* (pp. 135-150). Oxford, England: John Wiley & Sons.

Moore, L., Roberts, C. & Tudor-Smith, C. (2001). School smoking policies and smoking prevalence among adolescents: multilevel analysis of cross-sectional data from Wales. *Tobacco Control, 10*, 117-123.

Moskowitz, A. J., Kuipers, B. J., & Kassirer, J. P. (1988). Dealing with uncertainty, risks, and tradeoffs in clinical decisions. A cognitive science approach. *Annals of Internal Medicine, 108*(3), 435-449.

Moss, P. A., Cole, N. S. & Khampalikit, C. (1982). A comparison of procedures to assist written language skills at grades 4, 7 and 10. *Journal of Educational Measurement, 19*, 37-47.

Murnaghan, D. A., Sihvonen, M., Leatherdale, S. T., & Kekki, P. (2007). The relationship between school-based smoking policies and prevention programs on smoking behavior among grade 12 students in Prince Edward Island: A multilevel analysis. *Preventive Medicine, 44*(4), 317-322.

Nisbett, R. E. & Wilson,T. D. (1977). Telling more than we know: Verbal reports as data. *Psychological Review, 84*, 231-259.

Nykiforuk, C. (2004). Municipal tobacco bylaws: Use of geographic information systems to explore relationships between local ETS policy and community characteristics. Unpublished dissertation.

Tompkins, N. O., Dino, G. A., Zedosky, L. K., Harman, M., & Shaler, G. (1999). A collaborative partnership to enhance school-based tobacco control policies in West Virginia. *American Journal of Preventive Medicine, 16*(3, Supplement 1), 29-34.

Olson, G., Duffy, S., & Mack, R. (Ed.). (1984). Thinking-out-loud as a method for studying real-time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research.* Hillsdale, N.J.: L. Erlbaum Associates.

Ordonez, A. (2007). Prevalence of bullying among elementary school children as a function of comprehensiveness of anti-bullying policies and programs in the school. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 67*(8-A), pp. 2889.

Panter, A. T. (1990). A person by item model of responding to personality inventories. ProQuest Information & Learning). *Dissertation Abstracts International, 51*(2), 1020-1021.

Phelps, S. W. (1990). Identifying nonverbal problem-solving strategies through the use of think aloud procedures. US: ProQuest Information & Learning.

Popham, S. M. (1996). A test of an information processing model of responding to personality test items. ProQuest Information & Learning). *Dissertation Abstracts International: Section B: The Sciences and Engineering, 56*(11), 6456-6456.

Pressley, M. & Afferbach, P. (1995). *Verbal protocols of reading: the nature of constructively responsive reading.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Prior, L. (2003). In Silverman D. (Ed.), *Using documents in social research* (1st ed.). Thousand Oaks, CA: Sage Publications.

Rapley, T. (2007.). In Flick U. (Ed.), *Doing conversation, discourse and document analysis*. (1st. Ed.). Los Angeles, CA: Sage Publications.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Sacks, H. (1995). *Lectures on conversation*, Vols 1 & 2, Ed. Gail Jefferson. Oxford: Blackwell.

Shapiro, M. A. (1994). Think-aloud and thought-list procedures in investigating mental processes. In A. Lang (Ed.), *Measuring psychological responses to media messages*. (pp. 1-14). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Siegel, D. F. (1990). The literacy press: A process model for reading development. *Journal of Educational Research, 83*(6), 336-47.

Sienot, M. (1997). Pretesting web sites: A comparison between the plus-minus method and the think-aloud method for the world wide web. *Journal of Business and Technical Communication, 11*(4), 469.

Silverman, D. (2006.). *Interpreting qualitative data: Methods for analyzing, talk, text and interaction.* (3rd ed.). London: Sage Publications.

Stephens, Y. D., & English, G. (2002). A statewide school tobacco policy review: Process, results, and implications. *Journal of School Health, 72*(8), 334-338.

Strauss, A. and Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications.

Thompson, R. (2002). Reporting the results of computer-assisted analysis of qualitative research data. *Qualitative Social Research, 3*(2), 1-18.

Tompkins, N. O., Dino, G. A., Zedosky, L. K., Harman, M., & Shaler, G. (1999). A collaborative partnership to enhance school-based tobacco control policies in West Virginia. *American Journal of Preventive Medicine, 16*(3, Supplement 1), 29-34.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care: Journal of the International Society for Quality in Health Care, 19*(6), 349-357.

Van Den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology, 22*(5), 339-52.

Van Den Haak, M. J., De Jong, M. D. T., & Schellens, P. J. (2007). Evaluation of an informational web site: Three variants of the think-aloud method compared. *Technical Communication, 54*(1), 58-71.

van Someren, M. W., Barnard, Y. F. & Sandberg, J. A. C. (1994). *The think-aloud method: A practical guide to modelling cognitive processes*. Academic Press: London, UK.

Wedman, J., Wedman, J., & Folger, T. (1996). Analysis of analogical problem-solving processes via think-aloud protocols. *Journal of Research & Development in Education, 30*(1), 51-62.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Wertheimer, M. (1945). *Productive thinking*. New York: Harper & Row.

Wilhelm, J. D. (2001). Think-alouds: Boost reading comprehension. *Instructor, 111*, 26-34.

Yamauchi, K., Ono, Y., Baba, K., & Ikegami, N. (2001). The actual process of rating the global assessment of functioning scale. *Comprehensive Psychiatry, 42*(5), 403-409.

Yang, S. C. (2003). Reconceptualizing think-aloud methodology: Refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behavior, 19*(1), 95-115.

Young, K. A. (2005). Direct from the source: The value of 'think-aloud' data in understanding learning. *Journal of Educational Enquiry, 6*(1)

Young, S. (1982). Process models of reading: Some data on the initiation of processes. Accessed online July 2, 2009.

Zeisser, C., Lovato, C. Y., Zumbo, B.D., Pullman, A. & Manske, S. (2009). A Descriptive and Comparative Analysis of Canadian School Tobacco Control Policies. Poster presented at the National Conference on Tobacco or Health, Montréal QC.

Zumbo, B. D. (2005.). Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing. (Invited paper, Samuel J. Messick Memorial Lecture Award). Ottawa, Canada.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (pp. 45-79). The Netherlands: Elsevier Science B.V.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validity practice. In R. W. Lissitz, (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65-82). IAP- Information Age Publishing. Inc.: Charlotte, NC.

CHAPTER FOUR: CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

Review of the Purpose of the Dissertation

This dissertation examined in detail the cognitive processes of raters during tobacco policy rating through the lens of the substantive aspect of construct validity (Messick, 1994, 1995). In doing so, this research helped shed light on the validity of inferences from scores based on the Stephens & English (2002) tobacco policy coding rubric. Specifically, this dissertation showed how one can apply a think-aloud (TA) protocol in measurement to elucidate expert rater knowledge for a deeper understanding of the rating process itself. The purpose was to provide better understanding regarding the validity of inferences one can make from policy scores by providing contextualized and pragmatic explanations about the score generation process – along the lines of a new approach proposed by Zumbo (2009).

The dissertation was written in manuscript style according to the conventions outlined by the UBC Faculty of Graduate Studies. In the following, two related studies that are also prepared for submission as stand-alone manuscripts are summarized.

Summary of Findings in Light of Current Research in the Field

Study One Findings

Study one of this dissertation had two purposes: 1) to characterize school tobacco control polices from 10 Canadian provinces, describe key components of policies currently in place and provide interprovincial comparisons of key policy components, and 2) to use these policy characteristics as input into study two of this dissertation. Results from the characterization study indicate that school tobacco policies vary greatly in strength. An important message to the field based on the finding from this study was that the majority of school-developed policies currently in place in Canadian schools lacked the elements of smoking prevention education and smoking cessation access – components that are discussed in the literature as key policy elements (Pentz, Brannon, Charlin, Barrett, MacKinnon, & Flay, 1989; Wakefield, Chaloupka, & Kaufman, 2000). Clearly, board-developed policies had stronger components of smoking prevention education and smoking cessation access than school-developed policies. This result emphasizes the need for schools to focus more strongly on programs that reinforce prevention education and cessation when developing their tobacco control policies.

Study one findings strengthen current knowledge related to tobacco control indicating a need for much stronger emphasis on preventive actions reinforcing a smoke-free environment, e.g., through tobacco use prevention education (Murnaghan, Leatherdale, Sihvonen & Kekki, 2008; Pentz et al., 2008; Centres for Disease Control & Prevention, 1994). The policy characterization showed that in current school-developed policies there seems to be more emphasis on punitive actions and consequences after violations have occurred. This finding provides an opportunity for policy change in the future since current research (Pentz et al, 2008) has recognized smoking prevention education and cessation access as more effective than punishment in reducing youth smoking rates. Hence, these elements need to be more strongly addressed in Canadian school tobacco control policies, particularly in policies developed at the school level.

In the policy characterization study, board-developed policies also tended to be substantially longer than school–developed tobacco control policies. Findings regarding tobacco policy characteristics are comparable to those of existing research by Tompkins et al. (1999) in that board-developed policies (or county policies, in the West Virginia study by Tompkins et al.)

in general were more comprehensive than policies developed at the school level (e.g., addressing more key policy components).

The second part of study one, an interprovincial comparison of tobacco control policies, revealed substantial variation in the degree to which provinces addressed the key policy components. Comparing each province with respect to tobacco prevention education and smoking cessation access, the proportion of tobacco control policies addressing cessation was consistently higher than the proportion of policies addressing prevention. The exception was Newfoundland and Labrador, where both components were addressed by all policies. Across all other provinces, the majority of policies did not address tobacco use prevention education. In summary, Newfoundland and Labrador stood out as a leader, with 100% of its policies having a prevention education component, followed by British Columbia, Manitoba and Prince Edward Island. With respect to smoking cessation access, Newfoundland and Labrador was the only province with all school policies addressing this component, followed by New Brunswick, featuring more policies with the component than without; all other provinces had only a small proportion of policies that addressed smoking cessation. It was concluded in study one that this variation in policy strength may be related to strong provincial tobacco legislation that specifically applies to schools. These legislative acts support school boards in formulating tobacco control policies. Hence, local tobacco control regulations are commonly asserted to serve as statements of smoking acceptability in the respective community (Bonnie & Lynch, 1994; Brownson, Koffman, Novotny, Hughes & Eriksen, 1995).

The following school tobacco policy characteristics were identified for use in study two: policy type (school-developed or board-developed), policy length (long or short) and policy scope/comprehensiveness (minimal detail, medium detail and extensive detail).

The policy characterization study had several limitations. First, the results can only speak to the Canadian provinces included in the study; the Canadian territories (Yukon, Nunavut and Northwest Territories) were not represented in this study. In addition, the nature of study one was descriptive and exploratory and did not attempt to derive statistical inferences with respect to policy effectiveness. The range of policy components identified as key and described here is not exhaustive; other relevant elements could be identified, examined and compared. Future research should continue to monitor the school policy environment, particularly questions related to the implementation of policies.

Despite the above noted limitations, the results from study one can provide researchers and policy decision makers with valuable information regarding the status of school-based tobacco control policies affecting large numbers of students across Canada. In addition, results may be useful in guiding schools to update or strengthen their school tobacco control policies and focus on elements that need improvement, specifically the policy components of tobacco prevention education and smoking cessation access. Finally, the findings provide new information to tobacco researchers interested in instruments that can be used by schools for self-assessment purposes. The findings of the policy characterization study indicate that school tobacco control policies in Canada vary greatly in length and comprehensiveness, with only very few policies addressing smoking prevention or cessation programming. Schools are hence encouraged to develop more comprehensive policies that not only address prohibition but also enforcement, parental involvement and prevention and cessation programming.

Study Two Findings

Study two of this dissertation, the TA study of policy raters, investigated the substantive aspect of construct validity (Messick, 1994, 1995) in the specific context of school tobacco policy rating using a rubric based on Stephens & English (2002). Messick's notion of the substantive aspect of construct validity and his call for the study of process models were a useful foundation and motivation for examining this aspect in the context of policy ratings, particularly since this type of research approach had, to our knowledge, not been taken in the field of policy rating to date. It was necessary to investigate the process of rater responding and how this knowledge can help researchers interested in the validity of rating score inferences.

In addition, the TA study heeded the recent call by Cizek, Rosenberg & Koons (2008) and the Standards of Educational and Psychological Testing (AERA, APA, NCME, 1999) to study cognitive processes of responding as a rich source of construct validity evidence. The objective of the TA study was hence to examine processes of responding that generate raters' answers, obstacles raters encounter and how raters overcome them. The TA study presented three cognitive process models of policy rating, one for the policy rating process overall, one specifically for rating school-developed policies and one specifically for rating board-developed policies. This series of cognitive process models of policy rating helped answer the research questions of what cognitive processes are involved in the rating process, how these processes are organized, what obstacles distract raters during the rating process and how raters cope with obstacles.

Examination of the process models revealed that the cognitive processes and strategies expert raters use for written tobacco policies were of a wide variety and were used with great range of frequency. Key processes included assessing information completeness, formulating questions, making distinctions and generating inferences. As such, the cognitive strategies that raters used in the TA study are generally comparable to those strategies found by researchers of other cognitive processes, such as information processing and reading comprehension (Ericsson

& Simon, 1993; Olson, Duffy & Mack, 1984; Pressley & Afferbach, 1995). However, the raters' strategies also revealed cognitive processes very specific to tobacco policy rating. Examples are processes showing how raters used their expert knowledge and experience in managing obstacles to rating that stemmed from characteristics of the task environment, specifically involving policy characteristics.

Importance of the Linkage Between Studies One and Two

It is important to note that the results from the TA study were viewed in relation to the tobacco policy characterization study of this dissertation (Zeisser et al., 2009), since it provided the conceptual basis for the TA study. The combinations of rating items and rating objects (e.g., tobacco policies of varying types, lengths and comprehensiveness) formed a strong foundation for the validity arguments formulated in the TA study since they formed major components of the cognitive process models developed. As a part of this argument, it was important to focus on variations in rating processes when different policy types were rated, since this policy characteristic played such a predominant role in the rating process. This is the emphasis of the following section.

Variations in the Rating Processes as a Function of Policy Characteristics

The comparison of the overall model of rating with the models for school-developed policies and board-developed policies showed only minor differences and the overall model serves more or less as a general description of tobacco policy rating. However, the analyses that differentiated by policy type showed that there are subtle but interesting variations in the processes of rating school-developed policies, as opposed to board-developed polices. These variations speak to the rating objects – the policies themselves. While essentially the same rater behaviors were observed as main-and subcategories, the TAs about policies varied slightly in their frequency. The main process of ‘comparing to known standards’ occurred more frequently when rating board-developed polices; this can be understood as an interaction between rating and the nature of the object being rated. That is, board policies –being generally longer and more comprehensive – contain more statements that allow such comparisons. That this main process occurred with less frequency in the model for school-developed policies is an indication that here, raters are less able to make such comparisons due to lack of information. This finding is also in line with the observation that raters’ stating that key information they need to give a ‘yes’ rating is missing occurred 41 times when rating school-developed policies but only 31 times with board-developed polices. It was also an expected finding that the main process of ‘generalizing’ occurred more frequently for rating school-developed policies. School-developed policies tend to be much shorter and less comprehensive and thus, raters are more likely to use generalizations to come up with a rating response. However, for the same reason, it was surprising that the main process of ‘inferring’ occurred with almost the same frequencies for both policy types. Conversely, the process of ‘integrating multiple pieces of policy information’ was observed more often in the board-policy scenario, where raters frequently had to process and filter a very large amount of information available due to the greater comprehensiveness of this type of policy compared to school-developed policies.

The difference with respect to TAs pertaining to the ratings can likewise be explained by contextual variation in the task environment. When rating school-developed policies, raters used the sub-process of ‘double checking one’s ratings’, an indication of uncertainty about the rating decision made earlier. For board-developed polices, raters never used ‘double checking’; this is a subtle difference, but it can show that raters feel differing degrees of comfort and confidence with their decisions depending on the rating context – the characteristics of the object being rated. Paying close attention to the task environment as context of measurement is in keeping with the contextualized and pragmatic view of validity (Zumbo, 2009) and highlights the importance of context to explanation.

For meta-TAs, the main processes unique to rating board-developed policies – 'clarifying terms and meanings' and 'reflecting back on ones' understanding of policy content' can also be interpreted as a contextual difference in that raters' thinking is stimulated more when longer and more comprehensive policy content is available. In other words, the absence of these main processes when raters rate school-developed policies can be explained by the design difference and the resulting change of context: when the objects rated are school-developed policies, certain cognitive processes are not triggered or used by raters due to the characteristics of the policies. A similar explanation can be formulated based on a cognitive main category unique to rating school-developed policies: 'correcting oneself if contradicted' occurs only during rating policies with certain characteristics, and one possible explanation is that lack of clarity is more likely to be found in the shorter, less comprehensive school policies. With respect to meta-TAs, it was an expected result that raters more frequently 'stopped and wondered' or expressed a personal opinion when rating school-developed policies, but relied more on their expert knowledge when rating board-developed policies. When expressing a personal opinion, raters frequently referred to a lack of detail in information needed for rating school-developed polices, or complete absence of such information and their inability to rate this item with 'yes'. These findings from the models were substantiated in the detailed analysis of excerpts for rating obstacles and rater coping to overcome these, which will be reviewed in a subsequent paragraph.

A rather unexpected result was the use of the main process of 'clarifying terms and meanings' when rating board-developed policies and the absence of this process in the model for rating board-developed policies. This speaks to the observation that even though more comprehensive information is available to raters when rating board-developed policies, this information is not necessarily clear and succinct, so that raters need to employ a process of clarification before they move on to assigning a rating. Other key findings from the policy rating TA study pertain to the rating obstacles and how raters dealt with these.

Rating Obstacles and Raters' Coping Strategies

Once again, the context of what was being rated was utilized for clues with respect to rating obstacles and how raters coped with them. The TA study emphasized that while the most prevalent obstacle to rating either type of policy was that key information needed was partial or missing, there were important differences of rater coping, a finding which once again speaks to the importance of the rating context. In particular, board-developed policies appeared to be easier to rate then school-developed ones; while raters frequently critiqued both for lack of clarity and had to resort to deciding on partial information, raters managed by applying their expert knowledge when rating board-developed policies. This process did not occur in rating school-developed polices. Rather, for the latter, raters predominantly spent their coping efforts searching the policy for more information on which to base the rating on. This result indicates that board-developed policies - generally longer and more comprehensive documents - contained sufficient information to enable raters to overcome this obstacle by 'filling in the gaps'. In summary, the differentiation into types of policies and based thereon, separate content analyses were useful in understanding raters' cognitive processes and in developing process models of rating. Specifically, these differences in rating by type of policy show that raters indeed engage in different cognitive processes reflective of policy domain content. Hence, this finding speaks to the substantive aspect of construct validity (Messick, 1994, 1995), highlighting the identification of domain processes revealed in the assessment task – in this case, rating.

Implications for Research and Practice

Based on the findings from studies one and two, recommendations are made for schools, school boards and administrators to develop comprehensive school tobacco policies that provide sufficient information for purposes of policy assessments that tobacco researchers desire. It is also recommended that tobacco policy researchers communicate and collaborate more closely with the above mentioned stakeholders with the aim to create tobacco policies that are detailed, clear and comprehensive, thus serving schools and those interested in quantifying these policies alike.

With respect to the use of the Stephens & English coding rubric, it is emphasized that the higher frequency of 'no' ratings when rating shorter, less comprehensive school-developed policies is not a limitation of the rubric; it is in fact a reflection of the policy content domain. In light of the findings, one could recommend that schools develop longer, more comprehensive policies in the future, with the scope of the rubric in mind; however, such a recommendation needs to be made with the potential positive and negative consequences of assessment in mind (Messick, 1995). It is possible to create longer policy documents but without the desired content that is indeed necessary to achieve reduced smoking rates. In addition, a potential unintended negative consequence would be that measurement would guide policy development. That is, while a potential positive consequence is that schools indeed develop longer and more comprehensive tobacco policies, a potential negative consequence is that schools would be doing so simply in a 'teaching to the test' fashion.

Taking these cautionary notes with respect to future policy development into consideration, what is the implication for other researchers regarding the use of the Stephens & English coding rubric? The finding of different cognitive processes in expert raters when rating school-versus board-developed tobacco policies has highlighted two main points with respect to use of the rubric. First, this result demonstrated that raters indeed engaged in cognitive processes reflective of policy domain content available to them. Second, the lack of depth regarding TA responses when rating school-developed policies (e.g., predominantly brief responses indicating that the needed information to assign a 'yes' rating was missing) *cannot* be interpreted as a limitation of the assessment tool – the Stephens & English rubric. Instead, this result is due to actual policy domain content provided in the school-developed policy documents – the objects of rating. Therefore, the inferences one can make from the use of the Stephens & English coding rubric as a tool for measuring school tobacco policy strength are supported as justified, accurate and appropriate. Nevertheless, more insight into the process of rater responding can be gleaned with respect to rating board-developed policies specifically, since the TA responses from this type of policy were richer and contained more information about how raters went about assigning the scores. This richness in TA responses can be attributed to the larger scope of information provided in board-developed policies in general, in that it stimulated more thought processes in the experts while engaged in the assessment task.

Future Research Directions and Recommendations

With respect to school tobacco policy characterization, future research should continue to monitor and update the status of Canadian policies. Further, what is needed next is a closer collaboration among tobacco policy researchers, policy developers and school boards in order to reach a common understanding of what is meant when referring to 'strong school tobacco policies'. For example, a possible approach would be for collaborators and stakeholders in research and at the school board level to set forth clear guidelines on minimal school tobacco policy content. Simpler yet, since the existing Stephens & English (2002) coding rubric can serve as such a guideline, it should be disseminated to those who develop school tobacco policies.

Next steps for research should also focus on testing the validity of cognitive process models of ratings. Rather than accepting these models at face value, the models are but a first step, a foundation for future research. For example, the degree to which score meaning and action implications hold across respondents and populations as well as across settings and contexts is a persistent empirical issue for researchers. Hence, the cognitive process models presented here for different types of policies are a foundational first step; they should be replicated and empirically tested, ideally with a larger number of trained raters and more policies. In addition, the hope is that future researchers will be inspired to include an examination of qualitative TA data in their explorations of policies, either through the use of content analysis of TA protocols developing cognitive process models of rater responding or through expert interviews, to more completely capture the meaning of raters' decisions. Future researchers are also encouraged to consider a variety of coding schemes at either a more micro or more macro level than presented here, for examining participant-generated justifications for ratings and decisions. A continued research focus on the thought processes of raters that contribute to policy-related decisions can advance our understanding of the validity and fairness of these decisions. Such a focus would also add to the newer explanation-oriented view of validity and validation (Zumbo, 2009), as it would enhance understanding of the processes involved in rating score generation.

To summarize, based on the cognitive process models of tobacco policy ratings from chapter three, future research should focus on refining and replicating these models. To this end, the next step would be the construction of statistical models with the policy score as the dependent variable. One could then determine the statistically best fitted tree model to predict the policy score, using tree-based model analysis or a technique called Automatic Interaction Detector (AID), a type of decision tree technique. It can be used for prediction or for detection of interactions between variables. The next section will summarize the novel contributions of this dissertation.

Contributions of the Dissertation

Results from this dissertation make novel contributions to the fields of assessment and tobacco policy research in several ways. Study one was the first study to systematically examine and characterize a comprehensive set of school tobacco policies in Canada. Hence, study one provides a basis for policy input and revision with respect to what policy components need to be more prevalent in Canadian schools. As such, study one also provides impetus for the generation of more comprehensive school tobacco policy documentation that in turn yields richer input into the endeavours of tobacco researchers labouring to predict smoking outcomes from policies.

Study two was, to our knowledge, the first research to examine tobacco policy rating scores via the substantive aspect of construct validity and to provide explanation-based understanding of tobacco policy score meaning via the TA method. First, the results illuminate some of the cognitive processes underlying the responses of raters when they code tobacco polices using the Stephens & English (2002) coding scheme. The use of the TA method for illuminating expert rater knowledge is also an important contribution to the general discussion about the validity of inferences from rating scores; insights about how experts perform their task enables researchers working with the scores with respect to what can be said about these scores – in any field, not just in tobacco research. Specifically, the findings from study two focused on cognitive main-and sub-processes in different contexts of rating, allowing for the development of three separate cognitive process models of tobacco policy rating for maximum explanation. As a result, a deeper understanding of the response processes in generating tobacco policy rating was gained. This information allows for the gathering of validity evidence via the explanation of rating variations. The TA method was applied in a new way as a means to gather validity evidence and build a foundation for constructing theories of rating processes. Hence, the findings and proposed models could form a basis from where to develop substantive theories of cognitive processes in policy rating.

The second novel contribution of this dissertation is that results regarding cognitive categories of rating contribute to gathering validity evidence to strengthen inferences from the policy rating process in practical applications. This was accomplished by relating the cognitive processes of responding to the objects rated – written tobacco policies - , and their various facets in a systematic way, thereby improving the measurement quality (Zumbo, 2005) of the tobacco policy rating process. This issue has not been studied before, but should be researched according to Cizek et al. (2008), who warn that only a very small fraction of social science reports make use of the cognitive processes of respondents to better understand the scores so produced. The practical contribution of the TA study results to assessment is emphasized when constructive input into policy recommendations is made by drawing on raters' cognitive processes. Greater understanding of how the tobacco policy ratings were generated inspires greater trust in those ratings. Therefore, the ratings provide a more useful input for those aiming to improve pragmatic policy decision making.

The third novel contribution of this dissertation is that the insights into the rating process enhance the application of the quantitative policy indicators constructed from the policy ratings based on the Stephens & English (2002) rubric. In particular, the intended future use for these indicators is within a readily accessible tool for schools to evaluate their own smoking polices. This is a research goal in the context of larger projects in the field of tobacco control (e.g., at the University of British Columbia and the University of Waterloo). To accomplish this research goal, tobacco researchers at the University of British Columbia and the University of Waterloo desire information about the trustworthiness of policy scores generated using the Stephens & English (2002) rubric. The cognitive models of rater responding can be used as a foundation to form new hypotheses and research questions in the area of quantitative policy assessment in tobacco research, and to create stronger policy indicators.

The fourth novel contribution of this dissertation pertains to future rater training in tobacco policy rating contexts specifically, but also in other rating contexts generally. Many of the cognitive processes discovered in this dissertation are teachable in contexts other than tobacco. Therefore, the results can potentially help inform how future rater training is conducted and what type of training protocol will be developed, for use in tobacco policy- or other general rating contexts.

Finally, this dissertation makes a contribution to interdisciplinary research in general by applying methods traditionally used in education and psychology - such as the TA method and protocol analysis - to a research project at the intersection of applied population health sciences, policy development and psychometrics. The findings have implications beyond tobacco research; they are of relevance for policy rating in general, for example, rating environmental, social or health policies. The area of tobacco research is a good example of how one can apply the TA method to illuminate rater responding and cognitive processes in many other fields. This would be useful since it would help create a vibrant research exchange across disciplines. Policy characteristics such as scope and comprehensiveness – to name but a few of many possible variables - can now become variables in policy evaluation studies and policy research in general.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *The Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Bonnie, R. J., & Lynch, B. S. (1994). Time to up the ante in the war on smoking. *Issues in Science & Technology, 11*(1), 33-37.

Brownson, R. C., Koffman, D. M., Novotny, T. E., Hughes, R. G., & Eriksen, M. P. (1995). Environmental and policy interventions to control tobacco use and prevent cardiovascular disease. *Health Education Quarterly, 22*(4), 478-498.

Centers for Disease Control and Prevention (1994). Guidelines for School Health Programs to Prevent Tobacco Use and Addiction. *MMWR Morb Mortal Wkly Rep. 43*(RR-2).

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (rev. Ed.).* Cambridge, MA: The MIT Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). *Standards of validity and validity of standards in performance assessment.* Educational Testing Services, Princeton, NJ.

Murnaghan, D. A., Sihvonen, M., Leatherdale, S. T., & Kekki, P. (2008). The relationship between school-based smoking policies and prevention programs on smoking behavior among grade 12 students in Prince Edward Island: A multilevel analysis. *Preventive Medicine, 44*(4), 317-322.

Tompkins, N., Dino, G. A., Zedosky, L. K., Harman, M., & Shaler, G. (1999). A collaborative partnership to enhance school-based tobacco control policies in West Virginia. *American Journal of Preventive Medicine, 16*(3, Supplement 1), 29-34.

Olson, G., Duffy, S., & Mack, R. (Ed.). (1984). Thinking-out-loud as a method for studying real-time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research.* Hillsdale, N.J.: L. Erlbaum Associates.

Pentz, M. A., Brannon, B. R., Charlin, V. L., Barrett, E. J., MacKinnon, D. P., & Flay, B. R. (1989). The power of policy: The relationship of smoking policy to adolescent smoking. *American Journal of Public Health, 79*(7), 857-862.

Pressley, M. & Afferbach, P. (1995). *Verbal protocols of reading: the nature of constructively responsive reading*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Stephens, Y. D., & English, G. (2002). A statewide school tobacco policy review: Process, results, and implications. *Journal of School Health, 72*(8), 334-338.

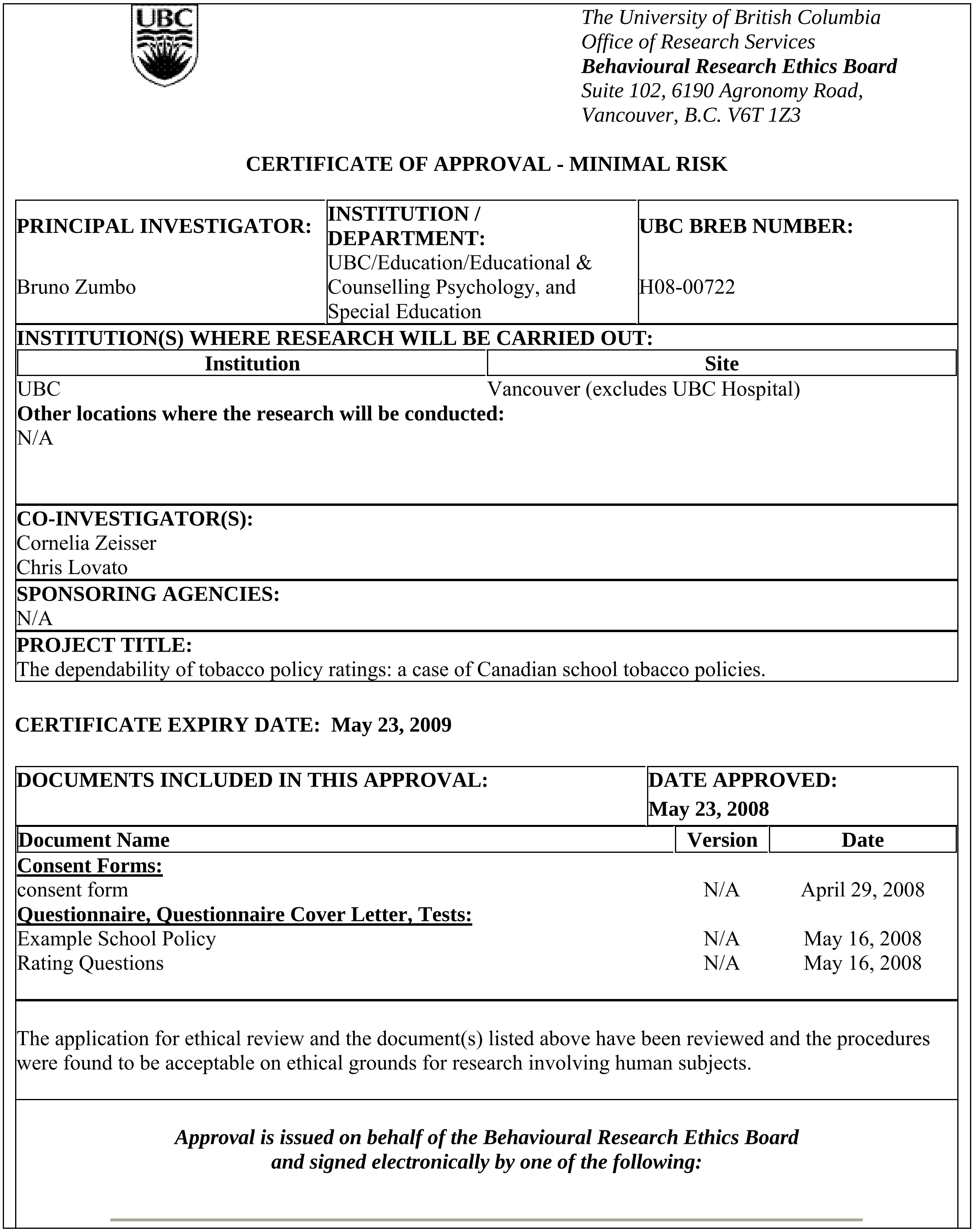
Wakefield, M.A., Chaloupka, F.J. & Kaufman, N.J. (2000). Effect of restrictions on smoking at home, at school, and in public places on teenage smoking: Cross sectional study. *British Medical Journal, 321*, 333-337.

Zeisser, C., Lovato, C. Y., Zumbo, B.D., Pullman, A. & Manske, S. (2009). A Descriptive and Comparative Analysis of Canadian School Tobacco Control Policies. Poster presented at the National Conference on Tobacco or Health, Montréal QC.

Zumbo, B. D. (2005.). Reflections on validity at the intersection of psychometrics, scaling, philosophy of inquiry, and language testing. (Invited paper, Samuel J. Messick Memorial Lecture Award.) Ottawa, Canada.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validity practice. In R. W. Lissitz, (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65-82). IAP- Information Age Publishing. Inc.: Charlotte, NC.

APPENDIX A



The University of British Columbia
Office of Research Services
Behavioural Research Ethics Board
Suite 102, 6190 Agronomy Road,
Vancouver, B.C. V6T 1Z3

CERTIFICATE OF APPROVAL - MINIMAL RISK

PRINCIPAL INVESTIGATOR:	**INSTITUTION / DEPARTMENT:**	**UBC BREB NUMBER:**
Bruno Zumbo	UBC/Education/Educational & Counselling Psychology, and Special Education	H08-00722

INSTITUTION(S) WHERE RESEARCH WILL BE CARRIED OUT:

Institution	Site
UBC	Vancouver (excludes UBC Hospital)

Other locations where the research will be conducted:
N/A

CO-INVESTIGATOR(S):
Cornelia Zeisser
Chris Lovato

SPONSORING AGENCIES:
N/A

PROJECT TITLE:
The dependability of tobacco policy ratings: a case of Canadian school tobacco policies.

CERTIFICATE EXPIRY DATE: May 23, 2009

DOCUMENTS INCLUDED IN THIS APPROVAL: **DATE APPROVED: May 23, 2008**

Document Name	Version	Date
Consent Forms:		
consent form	N/A	April 29, 2008
Questionnaire, Questionnaire Cover Letter, Tests:		
Example School Policy	N/A	May 16, 2008
Rating Questions	N/A	May 16, 2008

The application for ethical review and the document(s) listed above have been reviewed and the procedures were found to be acceptable on ethical grounds for research involving human subjects.

Approval is issued on behalf of the Behavioural Research Ethics Board
and signed electronically by one of the following: